A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated

**Module Leader and Supervisor 1:**

**Zia Ush Shamszaman | PhD**

Sr. Lecturer (Computer Science)

Department of Computing and Games

**Supervisor 2:**

**Dr Qiang Guo | PhD**

Lecturer (Computer Science)

Department of Computing and Games

**BY:**

**ABDUL AZIZ JAFFREE SHAIK**

W9326498

DATA SCIENCE

Traditional office vs Work from home or is there any Hybrid solution?

ACKNOWLEDGEMENT

I have worked hard towards this project. However, this would not be possible without the generous support and guidance of many individuals and organisation. I'd like to express my heartfelt gratitude to each of them who helped me  to work hard on this project. I'd like to thank them all from the bottom of my heart.

I am grateful to God for providing me with everything I needed to finish this project.

I am totally indebted to my course leader **Dr. Zia Ush Shamszaman** for his immense help, guidance and complete supervision on my work and providing me all necessary inputs and information the I needed in my project and for his guidance in finishing this project.

I'd like to thank my parents for everything they have done for me and their kind support and cooperation which helped me in the completion of this project.

I would like to express my special gratitude and thanks to university of **Teesside university** for giving me such facilities for attention, time and research.

My sincere appreciations and thanks also go to my inmates **Mr Vishwas, Mr Sarma, Mr Prashanth and Mr Arun** to find appropriate development and research towards the project.

Finally, I would like to thank all of them who have helped me willingly especially **Dr. Julien Cordry and Dr. Qiang Guo.**

Contents

[Abstract 4](#_Toc91084456)

[Introduction 5](#_Toc91084457)

[Project Description 5](#_Toc91084458)

[Relevant Theory 6](#_Toc91084459)

[Sentimental analysis 6](#_Toc91084460)

[VADER Sentiment 6](#_Toc91084461)

[Logistic Regression 7](#_Toc91084462)

[Project Specifications 8](#_Toc91084463)

[Scope 8](#_Toc91084464)

[Objectives 8](#_Toc91084465)

[Project requirements 8](#_Toc91084466)

[Tweet Scraping 8](#_Toc91084467)

[Getting access to the Twitter API 8](#_Toc91084468)

[Apply for a developer account with Twitter 9](#_Toc91084469)

[Get your Twitter API Key and Access Token 9](#_Toc91084470)

[Research Theory 10](#_Toc91084471)

[Research Question 10](#_Toc91084472)

[Research Objectives 10](#_Toc91084473)

[Hypothesis: 10](#_Toc91084474)

[Literature Work 11](#_Toc91084475)

[Database Research: 11](#_Toc91084476)

[Teesside Discovery 11](#_Toc91084477)

[ACM library 12](#_Toc91084478)

[Google Scholar 13](#_Toc91084479)

[Inclusion/Exclusion Process: 14](#_Toc91084480)

[Architecture 15](#_Toc91084481)

[Project Work Breakdown structure (Outline) 15](#_Toc91084482)

[Tweet Scraper Architecture 16](#_Toc91084483)

[Machine Learning Architecture 16](#_Toc91084484)

[Sentimental Analysis Architecture 16](#_Toc91084485)

[Logistic Regression Architecture 16](#_Toc91084486)

[Project Work Breakdown structure (Implementation) 17](#_Toc91084487)

[Datasets 18](#_Toc91084488)

[Consolidated Survey 18](#_Toc91084489)

[WFH VS WFO dataset 19](#_Toc91084490)

[Tweets Data 21](#_Toc91084491)

[Sentiment Dataset 22](#_Toc91084492)

[Emoji Dataset 22](#_Toc91084493)

[Methodology 24](#_Toc91084494)

[Data Extraction 24](#_Toc91084495)

[Consolidated Survey 26](#_Toc91084496)

[WFH VS WFO 27](#_Toc91084497)

[Twitter Raw Data 33](#_Toc91084498)

[Sentiments Analysis with Emoji: 41](#_Toc91084499)

[Dashboard 48](#_Toc91084500)

# Abstract

The Coronavirus pandemic has affected the normal course of life. Organizations all around the world were challenged and tested by tossing the traditional working model at office to question. Surprisingly, work from home model succeeded beyond imagination with accelerated adoption of remote working.

While the potential benefits of Work from home are substantial, finding the right balance in mixing virtual and on-site working would be very difficult in the post pandemic world, especially when the employees are not tied up forcibly in their houses, anymore. Organizational norms not framed well, can lead to erosion of social cohesion and sense of common culture, which can deplete & break the organization over a long period of time.

People all around the globe have shown their views and opinions using social media and expressed their general emotions regarding the pandemic that has taken all over the world as an unseen storm. Social media networking sites like Twitter, Facebook, and Instagram have shown an unprecedented increase in posts, tweets, and trends related to the pandemic coronavirus without any wait in a very short span. This paper showcases the sentiment analysis of Twitter trends and tweets related to work from Home (WFO) during coronavirus and how the sentiment of people and their expressed views on Twitter and how it affected over time. Furthermore, to determine the impact of various sectors and factors that have been affected on people’s daily aspects of life, tweets related to WFH, and online trends have been scraped and the sentiments of people over time were observed. To determine this various machine learning models such as logistic regression and Vader-lexicon were implemented for sentiment classification and their accuracies were determined.

# Introduction

Pandemic has had a significant impact on people’s lives around the globe. The outbreak has not only created chaos in people’s physical health, working conditions, economic impacts to name a few but also created a cranny in people’s minds around the world in various sectors. It has had a serious backlash on the psychological state of everyone which is more evident these days. Different platforms especially those of social media platforms like Twitter, Facebook, Reddit have become a great source to capture everyone’s opinions and emotions. During these times peoples have taken to social platforms to express their insights, fears, and opinions on the global pandemic. In this paper we are going to take an effort to analyse the tweets and trends of people’s sentiments and their changes during the pandemic globally. Due to lockdowns that were imposed by the governments across different countries because of the increase I spread of the virus, people were literally forced to opt and adapt to work from home and meetings, educational institutions must go with online learning platforms.

## Project Description

All over the globe people’s working conditions has been a major factor due to which some people show their regard as an inconvenience towards WFH due to lack of requirements such as high-speed internet or smart devices, while others regard it as a great option. This paper shows how different factory has affected WFH and how some daily aspects of life have been affected. This is shown by analysing their sentiments with respect to WFH scenarios. To ascertain or put a figure on sentiments on the tweets, a machine learning technique called sentimental analysis has been used to label and differentiate the tweets into Positives, Negatives and Neutral means. Two datasets were used to implement the analysis one was created from scraping tweets using “tweetscraper” and it was labelled the sentiments using sentiment lexicon. VADER was used then to classify the models. While sentimental analysis on the tweets can be reflected on opinions of people, it might not completely convey the exact impact. Therefore, we used the second dataset and we have done exploratory data analysis on it to showcase the various factors and conditions affecting people.

In this project we are going to use different Data science fields which includes Data mining to build data and forming a dataset using tweet scrappers, different websites to find and search the relevant dataset and building relations with the scraped data, we are using data analytics to analyse the data using various machine learning techniques and finally using Power Bi and Bigdata we are creating a dashboard to showcase the final work.

## Relevant Theory

All the techniques that we are using in this project have their own individual and specific usage and are the best techniques as per my knowledge to work on this specific topic.

### Sentimental analysis

A picture containing icon

Description automatically generatedSentimental analysis (also known as opinion mining) comes under a technique called natural language processing (NLP). It is generally used to determine whether the given input data is either positive, negative, or neutral. Sentimental analysis is frequently worked on textual type data to assist organisations in keeping track on their products and brand by taking consumer feedback and understanding their demands.

We have different types of sentimental analysis which includes aspect-based sentimental analysis, emotion-based detection analysis and finally fine-grained analysis. In most of the binary classification, the possible classifications are either positives or negatives. But when it comes to fine-grained sentimental classification, there are almost five different group- Very Positive, Positive, Negative, Very Negative. We have used binary sentiment and emotion-based detection.

### VADER Sentiment

Vader is a rule-based sentiment model which compares and analyses the effectiveness of words into 11 different benchmarks that includes Affective Norms of English words, general Inquirer, Linguistic Inquiry, Senti WordNet, including Word Count and other ML techniques that usually rely in SVM algorithms, Max Entropy and Naïve Bayes. This study describes the development. evaluation and validation of VADER.

Vader sentiment basically uses lexicon values or features(i.e., Word) which usually are labelled as either positives or negatives in accordance with their semantic orientations, then these orientations will be calculated as text sentiments. Vader sentiments returns the final probabilities of the given input sentences as positives, negatives, and neutrals.

Examples:

**“The food was great!”**

**Positive: 99% ; Negative: 1% ; Neutral: 0%**

**“The food was not bad”**

**Positive:40% ; Negative: 40%; Neural: 10%**

**“The food Was literally Worst”**

**Positive: 0% ; Negative: 99% ; Neutral: 0%**

**All the three responses will add up to 100%**

**Diagram

Description automatically generated**

### Logistic Regression

Logistic regression is basically a ML classifier algorithm which is generally used to predict the categorical dependent variable probability. Usually in logistic regression the dependent variable is only in binary variable format that usually contains data which is coded as either 0(Fail, NO, etc) or 1(Pass, Success, YES, etc.). In mathematical terms LG model predicts “P(y=1)=x” here, P(y=1) as a fn of x or x has a function as p(y=1).

Chart, line chart

Description automatically generated

Logistic regression is similar to other models like linear regression model, but LG utilizes cost function called “Sigmoid function” which is also known as logistic function. We have three types of LG functions Binary, multinomial and ordinal logistic regression and in this project, we will be using multinomial and ordinal logic to test the accuracy.

# Project Specifications

## Scope

While the potential benefits of Work from home are substantial, finding the right balance in mixing virtual and on-site working would be very difficult in the post pandemic world, especially when the employees are not tied up forcibly in their houses, anymore. Organizational norms not framed well, can lead to erosion of social cohesion and sense of common culture, which can deplete & break the organization over a long period of time. With this project we can analyse how covid-19 has affected people’s way of working and we can find out which kind of working system people would like to opt for future use.

## Objectives

The main objective of the project is to analyse:

1. People’s perspective’s, views, and sentiments on WFH, WFO and Hybrid system.
2. Analysis of WFO based on different sectors and factors.
3. If there is any impact of sudden change of working style on people with respect to different personal factors.
4. Finding out trends on social media platform and identifying which type of work system suits people after pandemic caused due to novel virus covid-19.

# Project requirements

## Tweet Scraping

We need Twitter API to gather or scrape the tweets data from Twitter. Twitter API is used to “Learn, engage with conversions on Twitter and Analyse programmatically”.

### Getting access to the Twitter API

You need API key and API access Token to make any kind of request to Twitter API in python or anywhere. For that we need to apply for a developer account on twitter and get approved. Once it is approved, we can create any project and associate with new or sample applications. Then the built sample app provides us with API key and its Access Token when is used to authenticate.

### Apply for a developer account with Twitter

* Firstly, open Twitter to [apply for access page](https://developer.twitter.com/en/apply-for-access) and apply for a developer account.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

* Fill up the application and answer all the questions. Then Review the Dev agreement and Policy and finally verify your email address. Once it is done you will successfully have twitter dev account.

### Get your Twitter API Key and Access Token

Once you get access to developer account you will be navigated to [Twitter Dev Platform](https://developer.twitter.com/en/portal/register/welcome).

Graphical user interface, application

Description automatically generated

Using the API key and key secret we can use it on any platform to access data from twitter API.

Graphical user interface, text, application, chat or text message, Teams

Description automatically generatedA picture containing graphical user interface

Description automatically generated

# Research Theory

This research project emphasizes to develop a sentimental analysis-based model for Work from home using various data analytics techniques. Using this research, we can contribute the modern working system to find out which system suits better for employees post covid-19 and opt for the best working way In the future. The research questions and research objectives are as follows.

## Research Question

Question 1: “*To what extent can aspect based sentimental analysis of Work from home (Twitter based) conducted using supervised machine learning techniques ( Logistic regression, Support Vector Machine, Linear Discriminant Analysis, K Neighbours Classifier, Decision Tree Classifier, Gaussian Naïve Bayes ,Random Forest Classifier, Bagging Classifier, Gradient Boosting Classifier) will be used to meaningfully find the accurate insights and to enhance the adaptability for the employees”. Sub-RQ: Will the analysis of sentiments(tweets and reviews) based on Work from Home can be used to support employers what way of working should they decide?*

Question 2 *: “Which sectors have been effected due to work from home and what each sector point of view says about future?”*

Question 3*: “How people with different personal factors affect the working style?”*

## Research Objectives

The specified objectives are basically a solution to aforementioned research question: **Obj A** was to scrap online views from different sources using several tools and programming techniques and then pre-processing them. After grabbing the data **Obj B** was to calculate the sentimental score for the data using natural language processing tools. **Obj C** is to find aspects for the surveys ( Work from Home) using NLP methods. Once we found the aspects **Obj D** was to deploy, inspect and evaluate the outcome of the prediction models using supervised ML techniques.

## Hypothesis:

Consider, H as hypothesis and X as work from home surveys, α is Input, β is output and µ is sentiments, ∏ is Aspects then H: α gives β where β¡=X,µ,α.

# Literature Work

## Database Research:

### Teesside Discovery

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | **Search Keywords** | **Applied Filters** | **Results Count** | **Assessment** |
| 1 | Work from home vs Office | None | 222,983Research Articles | Too many articles. Have extract recent papers, thus timeline is required |
| 2 | Issue of work from home compared to traditional office | None | 50,322 research articles | Broad domain: specific keywords are required |
| 3 | Work from home during COVID-19 | None | 42,785 research articles | Low technical articles |
| 4 | Work from home during COVID-19 | 2019-2021 | 42,749 research articles | Added “Filter” keyword; to find out recent journals on it but no change |
| 5 | Work from home vs office during COVID-19 | 2010-2020 | 1,506 research articles | Most of the journals shown were not related to the topic |

### ACM library

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | **Search Keywords** | **Applied Filters** | **Results Count** | **Assessment** |
| 1 | Work from home vs Office | None | 193,254Research Articles | Too many articles. Have extract recent papers, thus timeline is required |
| 2 | Issue of work from home compared to traditional office | None | 443 research articles | Broad domain: specific keywords are required |
| 3 | Work from home during COVID-19 | None | 1 research articles | No articles found |
| 4 | Work from home during COVID-19 | 2010-2020 | 1 research articles | Added “Filter” keyword; to find out recent journals on it but no change |
| 5 | Work from home vs office during COVID-19 | 2010-2020 | 0 research articles | No articles or research papers |

### Google Scholar

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | **Search Keywords** | **Applied Filters** | **Results Count** | **Assessment** |
| 1 | Work from home vs Office | None | 38,900 Research Articles | Too many articles. Have extract recent papers, thus timeline is required |
| 2 | Issue of work from home compared to traditional office | None | 14,000 research articles | Broad domain: specific keywords are required |
| 3 | Work from home during COVID-19 | None | 189 research articles | Few articles which are not completely related |
| 4 | Work from home during COVID-19 | 2010-2020 | 174 research articles | Added “Filter” keyword; to find out recent journals on it but no change |
| 5 | Work from home vs office during COVID-19 | 2010-2020 | 34 research articles | Not completely relatable. |

There were numerous research articles from various disciplines that informed this study, so Vader, LSTM, and ANN are all covered in this section of the literature review on sentiment analysis.

Most of the above results contains general articles but couldn’t find any specific research paper that studies about the surveys or statistical comparisons about these working ways and even the articles that are available doesn’t specifically compare both using any data, they actually express their own point of view on these working ways.

## Inclusion/Exclusion Process:

|  |  |  |
| --- | --- | --- |
| **Research Paper Title** | **Include/Exclude** | **Reasoning** |
| Work from home after the COVID-19 Outbreak |  | The paper does discuss the challenges associated with covid-19 and how work from home has changed this could be somewhat relevant to topic |
| Characterizing occupations that cannot work from home: a means to identify susceptible worker groups during the COVID-19 pandemic |  | The paper discusses issues of office system during covid; great insight information about all factors that might help intuitions to cope up |
| The impact of pandemic COVID-19 in workplace | × | Another paper with issues during pandemic but doesn’t concentrate on any of the working methods. |
| Work from home during COVID-19-disequilibrium of mental health and well-being among employees | × | The paper is off topic; thus, not relevant and to be used in the study. It’s basically studying about mental health of person |
| Work-from-home during COVID-19: Accounting for the care economy to build back better |  | Discusses about impact of covid on economy and advantage of work from home to build back financially. |
| A Comparative Study of Work from Home VS Work From Office: Preference of Women Employees in IT Industry | × | Have shown the problems and bias towards women in working industry but doesn’t show any clear comparison about work. |
| Employee engagement practices during COVID‐19 lockdown |  | Speaks about remote work and its uses but doesn’t compare anything |
| Global Sentiment Analysis of COVID-19 Tweets Over Time |  | This topic is pretty similar to the research I am doing but it concentrates only on WFH, but we are targeting specific country and looking for better solution. |

# Architecture

## Project Work Breakdown structure (Outline)

Diagram, schematic

Description automatically generatedThis project includes three major steps and each of them involves many internal tasks. Each of the major steps has an important relation with the oncoming processes.

## Tweet Scraper Architecture

Tweet scraper involves several steps that starts from twitter API and developer account to scrapping data into a csv or xlsx file.

Diagram

Description automatically generated

## Machine Learning Architecture

Machine learning was implemented using python Jupyter and we analysed the data using different algorithms and we have visualized different explorations using the scraper dataset and collaborative dataset created using GitHub. In the below flow diagram, we will see various steps involved during the analysis.

### Sentimental Analysis Architecture

Diagram

Description automatically generated

### Logistic Regression Architecture

Diagram

Description automatically generated

## Diagram Description automatically generatedProject Work Breakdown structure (Implementation)

# Datasets

In this paper we are considering three datasets. One comprises of consolidated surveys on Work from home with respect to covid-19. It includes various sectors and their opinions on work from home. The second dataset mainly focuses on factors effecting the way of working during pandemic and their opinions towards work from home and traditional office system. Finally, the third dataset is a scraped dataset that was scraped out from twitter trends and tweets especially on work from home with respect to the pandemic and how its trends have been shown on internet in last two years.

## Consolidated Survey

Table

Description automatically generatedHere we are considering a excel file “consolidated\_survey.xlsx” that was taken from [KDnuggets](https://www.kdnuggets.com/). It describes about how different sectors were affected from covid-19 pandemic and also shows surveys about people choices towards work from home and their different perspectives.

The above dataset which is available [here](https://docs.google.com/spreadsheets/d/1AiCFVX0905lnCUMjZdjsijGDh5v4ZFVa/edit?usp=sharing&ouid=107220898721675825407&rtpof=true&sd=true) in my folder and is an opensource folder. It consists of different attributes

* **ID** (represents the survey number)
* **Completion Time** (show when the survey form was finished by the person. This clearly shows what time of year the person has given his survey and we can check the sentiment of the person by the covid situation at that time)
* **Sector** (Sector shows which filed of work that person belongs to)
* **For how much duration you have done work from Home in months.** (Survey Question)
* **Do you have required infrastructure for working from home?** (Survey Question)
* **Do you feel work from home is better than working at office?** (Survey Question)
* **Overall satisfaction with Working from Home** (Overall satisfaction with Working from Home explains persons opinion and sentiment on work from home.)

## WFH VS WFO dataset

Here we are considering a CSV( Comma Separated Variable) file “WFH\_WFO\_dataset.csv” that was taken from [Kaggle](https://www.kaggle.com/anninasimon/predict-if-people-prefer-wfh-verses-wfo-data?select=WFH_WFO_dataset.csv). It describes about how different factors affect people while work from home during covid-19 pandemic and also shows different aspects about people and how they manage things.

A screenshot of a computer

Description automatically generated with medium confidence

The above dataset which is available [here](https://drive.google.com/file/d/1NSCIZLm6Eg3mRX8nANdd7D73oRG8-A6M/view?usp=sharing) in my folder and is an opensource folder. It consists of different attributes

* ID - Unique Identifier for each record
* Name ( User Name )
* Age ( Age of Person is main aspect as it defines the persons interest towards WFH)
* Occupation (Defines persons working sector and is useful to find out weather it suits WFH or not)
* Gender( Male / Female)
* Same\_ofiice\_home\_location ( Whether the person lives in office allotted location or not)

Answer : Do you reside in the base location as your office? --> yes or no

* Kids ( If the person has kids and we can find if its effects his working performance and active time)

Answer: if they have kids or not. --> yes or no --> yes or no

* RM\_save\_money (Does WFH is saving person money in terms of expenditure)

Answers the question :Do you feel you were able to save money with remote work? --> yes or no

* RM\_quality\_time (Could he utilise as much as time in good quality performance during WFH or not)

Answers the question : Do you feel that remote work has given you more quality time with family/friends? --> yes or no

* RM\_better\_sleep ( Could he have calm and happy sleep while WFH)

Answers the question : Has your sleep cycle improved with work from home?

* calmer\_stressed ( whether WFF is leading his mind into Stressed, or he is happy with WFH)

Answers the question : Are you calmer or more stressed than usual since remote work began?

* RM\_professional\_growth ( Is there any position hike while in working at home ) “Rating”.

Answers the question :On a scale of 1-5, Do you feel WFH has affected your professional growth adversely?(5-yes it's affecting me badly,1 No it doesn't affect me)

* RM\_lazy (Could he handle work intime or is he lazy during WFH when compared to in office work)

Answers the question : On a scale of 1-5,do you feel WFH has made you lazy?(5-extremely lazy,1-Nope, not lazy at all)

* RM\_productive ( Does WFH had impact on his Productivity) “Rating”.

Answers the question : On a scale of 1-5, are you more productive with working remotely?(1-not productive at all,5-extremely productive)

* digital\_connect\_sufficient (Does he has enough resources to connect digitally)

Answers the question : Do you feel digital connect is sufficient?

* RM\_better\_work\_life\_balance ( How good is he managing his personal and professionalism) “Rating”.

Answers the question : On a scale of 1-5,do you feel you have a better work-life balance with remote work?(1-No not at all,5-Yessssssss)

* RM\_improved\_skillset ( Could he improve his experience and performing skills) “Rating”.

Answers the question : On a scale of 1-5,how much has your skillset improved in the last two years?(skillset with respect to your work domain,1:Not improved,5:Improved drastically)

* RM\_job\_opportunities ( Has he got any other job opportunities or position during the time)

Answers the question : Do you think there are more job opportunities with remote work?

* Target ( Has he managed to fulfil the targets during WFH) “Rating”.

Answers the question : In the future which of the following do you think is more suited for you?(Sorry, but hybrid isn't an option here)

## Tweets Data

Here we are considering a excel file “Tweet\_Sample.xlsx” that was scraped from twitter using tweet scraper and it detailed version and step by step work was mentioned above. It describes all the trends that run on work from home during pandemic both tweets and re-tweets.

A screenshot of a computer

Description automatically generated with low confidence

The tweet samples dataset was built by grabbing tweets containing hashtags “#WFH”, “work from home” between the dates November 2019- September 2021. A subset of this corpus was manually annotated by me with 5 labels( Type, Source, text, Created\_AT, View). This dataset consists of more than 8000 tweets including all retweets, replies and commented tweets. This is used to evaluate polarity but can be used to valuate classifications too since it has neutral tweets in it too.

The above dataset which is available [here](https://docs.google.com/spreadsheets/d/1x-82yRyNnlyt69zTwOjZRxFmpBIrmNjF/edit?usp=sharing&ouid=107220898721675825407&rtpof=true&sd=true) in my folder and is an opensource folder. It consists of different attributes

* Type (Describes the tweet type either it is “Tweet/Retweet/Replied to”)
* Source ( Describes the source or device used such as Android/iPhone/iPad/Web/other)
* Text ( Original Tweet )
* Created\_AT ( When was the tweet done)

We use this dataset to analyse the sentiment analysis and find the final view on work from home.

## Sentiment Dataset

Table

Description automatically generated with medium confidenceThis dataset was created using the above dataset after splitting it into sentiments this is used to analyse the accuracy of sentiments and know how much absolute the tweets data was used to analyse.

The above dataset which is available [here](https://drive.google.com/file/d/1g9zEQlFVbi-hzoB8og8c7ttCJ64Cl-u-/view?usp=sharing) in my folder and is an opensource folder. It consists of different attribute

* Text ( shows the tweet )
* Positive ( Describes how much Percentage of Positive sentiment the text has)
* Negative (Describes how much Percentage of Negative sentiment the text has)
* Neutral ( Describes how much Percentage of Neural sentiment the text has)
* Compound
* Status ( Overall sentiment of the text)

## Emoji Dataset

Graphical user interface, application

Description automatically generatedWe use this dataset “ Emoji\_Sentiment\_Data\_v1.0.csv “ to read all the emotes that were used in “tweet\_samples.xlsx” instead of converting them into text format. Instead of removing the emotes this dataset really helps us to find out the sentiment of the data up to 80% accurate.

The above dataset which is available [here](https://drive.google.com/file/d/19wyANFLlCIdSrkf315k5_1RIgfB4AUN3/view?usp=sharing) in my folder and is an opensource folder. It consists of different attribute

* Emoji: contains various types of emoticons in it.
* Unicode codepoint: Every emoji has an international standard unique code. This column describes every emotes unique code.
* Occurrences: Describes how often it is usually used
* Position: Shows text position and how its text values occupy.
* Negative: value of negativity in the emote
* Neutral: value of neutrality in the emote
* Positive: value of positivity in the emote
* Unicode name: International standard name of the emote
* Unicode block: Specifies either it is a ‘emoticon’ or ‘miscellaneous symbol and photographs ’.

# Methodology

## Data Extraction

In order to grab tweet information from twitter we will use, and API called twitter API which is used to facilitate the idea behind grabbing the information. This Application programme interface will let us read and then write the tweets data and can be accessed to high volume tweets on a specific subject.

The primary step of this data extraction is to apply for a twitter dev account. This is mentioned above how to apply for a twitter dev account. Once the application is being reviewed and accepted, we can use the REST API to filter, search, access, and download the twitter data we require in this project.

Once account is approved and you are on API dashboard, you will be guided automatically about how to create an App and how to start a project. When its created it will generate two keys and Tokens. Then We need to use Python for data extraction using a library called “tweepy”.

Text

Description automatically generatedStep 1: We need to install the libraries “tweepy” and “preprocessor”. These are used to pre-process the tweet data written in python.

Step 2: In this step we are importing the libraries and giving our API access data to authenticate our developer account information that was created earlier and using this we can connect out Twitter API with out python and extract all the required data.

A screenshot of a computer

Description automatically generated with medium confidence

Step 3: Here we are specifying the timeframe, limit of tweets, language and how we need the data to be stored.



Text

Description automatically generated

Text

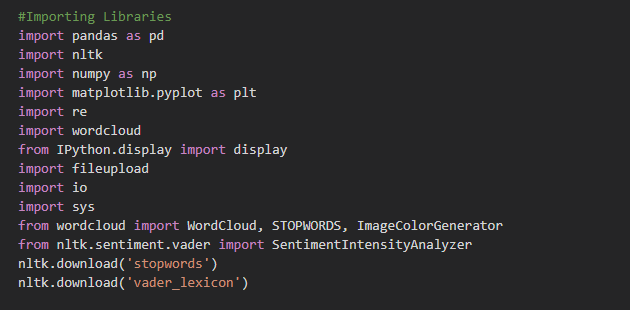
Description automatically generatedText

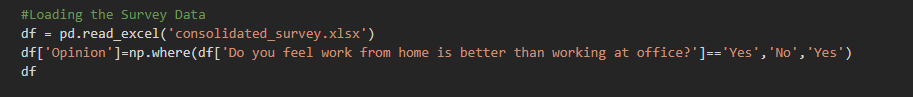
Description automatically generatedStep 4: We are going to comprehend all the specification we need to extract, and we have added all the filters. In the next line of code, we have specified how want it to be extracted and in which format it should be saved.

Step 5: Finally, the output is stored in raw data in comma separated variable file.

## Consolidated Survey

In this data we are going to find how Covid-19 has affected peoples working lifestyle and how different sectors behave with respect to WFH and how much percent of people have opted to go with WFH.

Step 1: Firstly, we need to import all the libraries that we might require in this process, and we rename them accordingly.

Step 2: Now we are going to import the consolidated survey file from the local directory.

A screenshot of a computer

Description automatically generated with medium confidenceStep 3: Now we are going to define specific columns and perform grouping and aggregation operation based on different industrial sector.

Chart, pie chart

Description automatically generatedOutput: We can see how each sector has shown their opinion on WFH and how it differs from each sector, and we can see sectors like Chemical and Electronics have shown less interest in WFH as they cannot work from home as most of their work is based on machines.

Text

Description automatically generatedStep 4:In this final step we are going to explore the final voting on work from home and overall choice of people on WFH based on the survey.

Output: We can see percentage of votes are almost half and half as most of the sector’s cant opt for work from home due to their working style.

Graphical user interface, chart, pie chart

Description automatically generated

## WFH VS WFO

This dataset was taken from Kaggle, and we are going to analyse how different factors affect people while work from home during covid-19 pandemic and also shows different aspects about people and how they manage things.

Using these same features, I have tried to see how much my model can predict. The target column has the right output with which we will train our model. We then test the model and compare the actual and predicted values.

Now let get into the working model and first we will import the dataset then we shall do basic exploratory data analysis on the data to find all the major findings and comparisons then we shall train the model to use classifier logistic regression.

Text

Description automatically generatedStep1: First step is to import all the libraries that we might require in this process, and we rename them accordingly. Here we are importing seaborn for plotting graphs and during this process we get lot of errors so we are going to import “warnings” to ignore these errors.

Text

Description automatically generatedStep2: This step involves importing the comma separate values file “WFH\_WFO\_dataset.csv” and once it is imported, we are going to check the data by using the function “data.head()” to check if the imported data is exact data or not.

Text

Description automatically generatedStep3: From this step we are going to do basic exploratory data analysis to find out different comparisons in the dataset. So here we are considering WFH vs WFO with this comparison we can find how much percentage of people prefer work from home and work form office this is compared with the target variable.

Chart, bar chart, treemap chart

Description automatically generatedOutput: Here we can see people prefer WFH more when compared to WFO we can check this with the count value with respect to target variable .

Chart, bar chart, treemap chart

Description automatically generatedChart, bar chart

Description automatically generatedStep4: Similar to the above function we have done more exploration on “ Male vs Female” , “ calmer vs stressed”, “Occupation based” and finally people with “kids vs no kids” to find out their behaviours with respect to their individual personal factors.

Chart, bar chart

Description automatically generatedChart, bar chart, treemap chart

Description automatically generated

Text

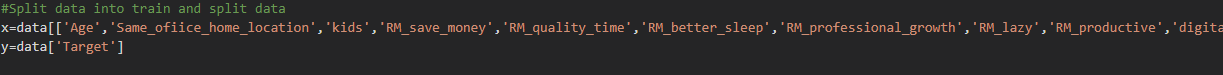
Description automatically generatedStep5: This step involved finding how much percentage of people want to opt which type of system for the future as a working style and we are comparing feedback value count to find out.

Chart, bar chart

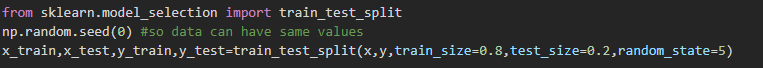
Description automatically generatedOutput: We could clearly see that people have opted for hybrid model more than WFH and WFO.

Text

Description automatically generatedStep6: From this step we will check the accuracy of the target variable and perform logistic regression classifier. Firstly, we are converting the object variables “Yes” and “No” into ‘1’ and ‘0’ values for all the attributes that has these variables.

Step7: Now we are considering all the attributes except target variable as “X” data and we are considering the attribute ‘Target’ as “Y” data.

Step8: Now from sklearn we need to import test and train to split the data.



Step9: During these steps we are train the model and using the lo formula . Now the next step involves prediction of data using test data and using this we can find accuracy .

Linear Regression: y =

In the above equation y is dependent variable and X1,X2…Xn is explanatory variable.

Sigmoid Function: p=

When we apply sigmoid function on linear regression

P=

So, the final logistic regression equation become

.

Logistic function is also known as sigmoid function usually gives a curve almost similar to curve with shape ‘S’. It can take any real valued number and can easily map the them into values that vary between 0 and 1. During this process if the curve moves towards positive infinity, them the predicted value of y will become 1, and if it moves towards negative infinity, then the predicted value of y will become 0.If the sigmoid function output is more than 0.5 then we can classify it as YES or 1, and if it is less than 0.5 then we can classify it as NO or 2. For example: If the output is 0.82, then we can say the probability of term as there is a 82% chance that the person wants work from home.

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceOutput: Using the matrix we could check the accuracy score and from this below output we can clearly see the accuracy approximately 88% and out of 24 values we predicted almost 21 correct values and 3 incorrect values.

Text

Description automatically generatedStep10: This is the final step; in this step we are going to build a confusion matrix image to calculate the accuracy percentage and we are going to show the correct to incorrect valued matrix

A picture containing graphical user interface

Description automatically generatedOutput: The output shows clearly that 13+8= 21 are correct values and 2+1=3 are incorrect values.

The overall gained accuracy by logistic regression is 88% which is high, and it clearly says that the total performance of the predicted model is as high because accuracy defines the high performance of model. The gained precision by logistic regression is 87% which is highly enough to contribute the prediction as positive for the model. We got 93% in recall which is very high and can support perfectively to positive predicted models.

## Twitter Raw Data

In this dataset the emoticons or any kind of symbols or scrips written in any other language except English will be shown as symbols in the dataset so we will be removing the during the filtering process. We are going to do this to check the sentimental value precession difference with and without emojis.

Text

Description automatically generatedStep1 : In this step we will be importing the dataset which in in excel format and we will be checking the data using the tweets.head() function.

Text

Description automatically generatedOutput: We can see the data has been imported successfully and we can see all the attributes were clearly mentioned.

Step2: In this step we will be checking the word frequency of the text and analyse which word or symbol is used majorly.

Text

Description automatically generated with medium confidenceText

Description automatically generated

Step3: Now in this step we have removed all the symbols, emoticon texts, numbers and other miscellaneous values.

Text

Description automatically generated

Text

Description automatically generatedStep4: In this step we will be generating the word count and we are assigning the figure sizes and other general things like colour of the background and axis etc.

Text

Description automatically generatedRemoving general words from the Wordcloud: We are going to remove frequent sentence adjectives and other general words from the sentences.

Text

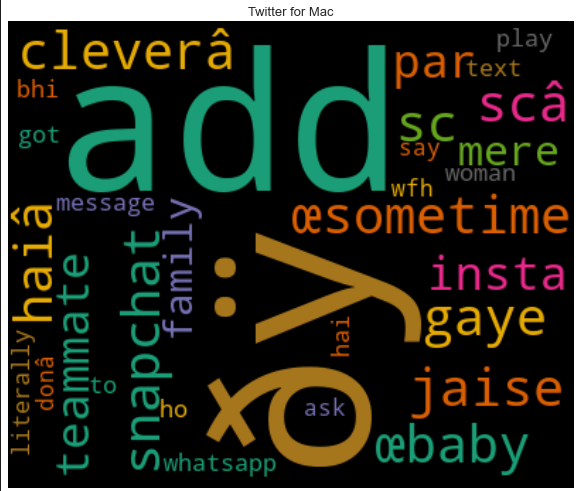
Description automatically generatedOutput: We generated the Wordcloud in twitter format by importing twitter png file and using that shape we created it in twitter shape.

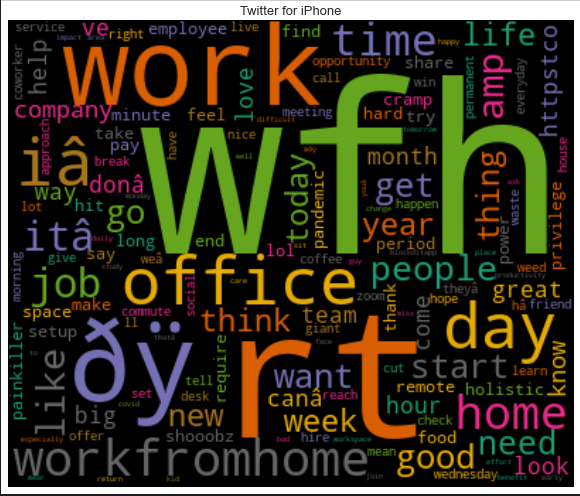
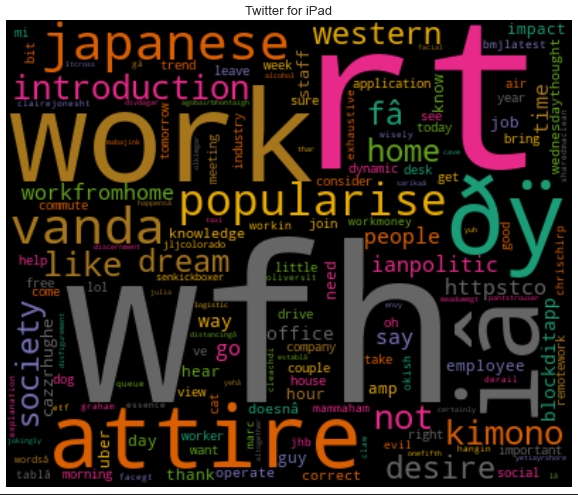
Step5: In this step we are going to build wordclouds for different sources like whether its tweeted or retweeted using iPhone, Android, Mac, Web, iPad and we have shown word frequency from each device and using that we can find out the most frequent words by each device.

Text

Description automatically generated

Output: Here are the wordclouds for each source





Step6: Now we are going to use VADER lexicon to spit the data into polarities (Positives and Negatives)

Output:



Text

Description automatically generated

Step7: We are going to classify the status of the text using the percentages of positives, negatives and neutrals by having the compound values Here we are assigning C >= 0 as negative else it is positive.

Text

Description automatically generatedOutput: We can see the status column with the additional terms neg or pos to define the text sentiment .

Step 8: We are going to use “openpyxl” to edit and export the above output as a dataset for further use and we are performing grouping and aggregation operation based on sentimental analysis to find the most votes on WFH.

A screenshot of a computer

Description automatically generated

Chart, pie chart

Description automatically generatedOutput: We can see the Percentage of employees saying “Yes” to returning office is just around 25% and the rest 75% don’t want to return office and wanted to opt for work from home and the votes percentage is quite higher for negative.

Text

Description automatically generatedChart, pie chart

Description automatically generatedText

Description automatically generatedStep9:In this we are trying to summarise by grouping for the consolidated data results and finding the number of votes.

Outputs:

Text

Description automatically generatedChart, line chart

Description automatically generatedStep10: Finding the distribution of polarities in the dataset.

Output:

Text

Description automatically generatedStep11: Using split function we made the graph much accurate and now the distribution of polarities vary compared to previous one.

Chart, line chart

Description automatically generatedOutput:

Chart, bar chart

Description automatically generatedText

Description automatically generatedStep12: Using the polarities we have deployed them as negative as work from Home, Positive as work from office and Neutral as Hybrid and we can see the final outputs.

Outputs: Most people want to go with Hybrid system when compared to WFH or WFO.

## Sentiments Analysis with Emoji:

In our previous work we have used excel sheet that has emotes converted into text, but in this analysis, we are going to use a separate emoji dataset to firstly train and them find the sentiments of the tweets exactly. Using emojis we can find almost exact sentiment of the text as emojis shows exact sentiment of a text.

Step1: We are going to import all necessary libraries that we are going to use during this process.

Text

Description automatically generated

Step2: Importing the actual tweet scrapped excel dataset “sample.xlsx”

Text

Description automatically generated

Output:

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generatedStep3:Converting the Status values “neg” as “0” and “Pos” as “1”

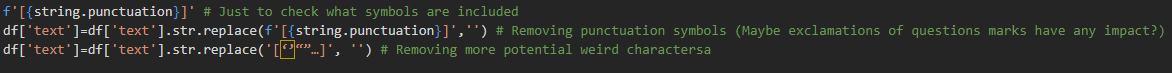
Output:

Text

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceStep4: In this step we are going to Import the emoji dataset that we are going to use to train.

Output:

Step5: We are checking the symbols and remove the unnecessary characters or symbols involved in text. Here we are trying to remove various objects such as punctuations symbols, weird chars, etc.

Step6: We are going to do general exploratory data analysis on the dataset. Here are we are checking which type of tweet has been done more times.

Output: We can see most of them are retweets.

Chart, bar chart

Description automatically generated

Step7:Understanding the Binary values with respect to True and False

Graphical user interface, application

Description automatically generatedOutput:

Step8: Here we are finding the percentage capital for the text.



Text

Description automatically generated with low confidenceOutput:

Step9: We are plotting the Positives and Negatives using distribution plot for marking between perc\_capital and density of Status variables.

Text

Description automatically generated

Output: We can see the density of negatives is higher than positives and the values and simultaneously reducing as the value of perc\_capital increases.

Chart, line chart, histogram

Description automatically generated

Step10: Striping the lower-case letters as capital may have higher prediction impacts



Text

Description automatically generatedOutput:

Step 11: Checking the character length and plotting the distribution chat for char\_length with respect to density using Status values.

Text

Description automatically generated

Output:

Chart, line chart, histogram

Description automatically generated

Step12: In this step we are going to generate the Wordcloud for most numbers of words used frequently as positives and negatives.

Text

Description automatically generatedOutput:

Text

Description automatically generatedStep13: Checking the accuracy using naïve bayes and we got trained accuracy as 77% and developed accuracy as 56% which is very less

Text

Description automatically generated

In order to increase the accuracy, we shall use logistic regression classifier.

Step14: We have used logistic regression classifier to check the sentiment accuracy and we got trained accuracy as 74% and we could produce the developed accuracy as 75% which is a good accuracy.

Text

Description automatically generated

Text

Description automatically generatedStep15: Now we are going to compare the accuracies using different classifiers such as logistic regression, Decision tree classifier, K nearest neighbour, Linear Discrimination analysis , Gaussian Naive bayes, Support vector machine, Random Forest, Gradient boosting, XGB, Bagging classifier. Here we are importing all these libraries.

Step16: Seeding all the models for cross validation

Text

Description automatically generated

Step17: Evaluating the model by testing accuracies for different classifiers and plotting a box plot to showcase the difference between each classifier.

Text

Description automatically generated

Text

Description automatically generatedOutput: We can see that SVM has highest with 72% and LDA has the lowest of all with 55%.

Boxplot:

Chart, box and whisker chart

Description automatically generated

# Dashboard

Using Power BI we are creating visuals using three different datasets “Consolidated\_survay”, “WFH vs WFO” and finally a new dataset that has different questionnaires. Using these datasets, we are going to show different graphical outcomes and we are going to plot them using various types of graphs

## Page 1: WFM Data

A picture containing diagram

Description automatically generatedUsing the dataset “WFH vs WFO” we have created a dashboard that has various explorations using different types of graphs and we have also shown all the major findings that we could analyse using power BI.

Using all the questions below we developed a Dashboard as shown in the above figure and we used Stacked Bar Graph, Bar graph, Pie Chart, Tree Map and Stacked column chart.

* How different sectors of work effect the sleep cycle of person?
* How much percent of people are digitally connected?
* Productivity of people with respect to their living location (Office houses or Private housing)
* People’s laziness with respect to their age
* Can Productivity and balanced life is possible?
* Other comparisons such as saving money, Occupation % etc

## Page 2: Survey

During this analysis we have used “Consolidated\_survay” dataset and we have answered different types of questions and managed to get proper plots and expected outcomes accordingly

* Which mode of work is mostly preferred?
* Which domain can easily opt to work from home?
* Do people feel work from home is better than working at office?
* How much % of people have Work from home infrastructure?
* For how much duration you have done work from Home in months?

We have created a dashboard answering all the question above and built it as shown in the below figure. This was done using Line graph, Pie chart and Bar graph.

## A picture containing graphical user interface Description automatically generated

# Discussion

Social media in these days can be considered as a big tool for any type of analysis and we are taking this as a major opportunity for us to do data analysis and to exploit it biggest areas. We can analyse people’s opinions, their interests, can judge how and predict impact on products and many other critical works can be done in terms of marketing strategies, understanding customers feedbacks and build new ideas and strategies to overcome them. We can expect that this research can be used by many companies or institutions in any field to benefit them.

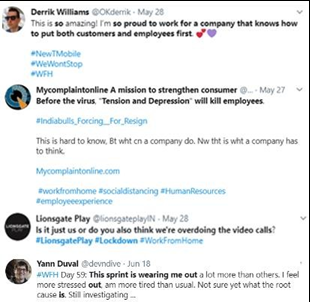
The implementation of sentimental analysis from the extracted tweets and including work from home as its mainstream search is never done before but still this concept has given us very interesting results which could be analysed using machine learning algorithms and also using NLP’s. During this process the dispersion results in all the output figures and the graphs shows us a clean and interesting outcome. The graph states that most of the tweets tend to have neutral and negative than positive polarity when it comes to work from office and that almost all the tweets dispersed tend to be an opinion based more than being a fact based. We can clearly say from the Pie graph that the most people are showing a negative emotion to go back to office after covid-19 and are having a positive behaviour towards working from home.

The Pie graph of tweet polarity clearly shows us that there is a high density in neutral tweets when we worked on raw data by converting the emoticons into text format. This result clearly tells us that we need to improve the way we are cleansing up the data and tweet texts. So, during the cleaning stage, we are separating out all the keywords that express the opinion of that tweet. Now we are using the initial scraped data that contains all the emoticons and on the other side we are training some part of data using emoticons dataset and we can see the after results that there is more density in the positives range of polarity when compared to negative tweets. The graph of both the formats can clearly show us the difference when we checked the accuracies for both of them. The tweet subjectivity also clearly explains us the outcomes if the tweet content is mainly an opinion based or a fact. This must have tended to show a higher number of negative tweets. When the tweet tends to be completely an opinion based by the user concerning to work from home or a topic or any kind of situation. It is totally correct that the specific sentimental analysis has more density in the opinion-based range.

Using the Wordcloud method, we can clearly know which words have higher frequency in the dataset. Words such as WFH, Work from home, Covid, people, time, job, employee, office, remote, hiring can define an important idea and the reason for the twitter sentiment. These words can clearly shoe us what is affection workers opinion towards wfh and how we can work in a strategy to comply and attract the major modern world problem. Office based setup in home and using it as a workplace will play a major and important role in employer productivity. Companies could provide their employees laptops, desks, remote, working accessories. Headphones and other necessary requisites that fulfil the workplace. The trends and hashtags can help us identify other missing data as they are major acronyms and alternate ways to name home workplace as a modem work from home. Other hashtags used are about remote work, office designs, covid, coworking, wfh meetings, Microsoft teams, zoom, google collab, wfh accessories and workspaces.

# Insights and Recommendations

All the insights we could discover by studying this topic using topic modelling, the hashtags, word clouds, and finally algorithms like logistic regression, sentimental analysis, all these helped us to brainstorm few suggestions.

Topics such as ‘WFH recommendations’, ‘WFH environment’, ‘Family’ and ‘Sector’ revels us both material and immaterial needs of people working from home.

Topics like these suggest us how people are experiencing various problems adjusting their personal lifestyle with WFH and how they have an impact on sleep cycle, kids due to overwork. These outcomes and findings are especially augmented by the words discovered in all the generated wordclouds such as WFH, Work from home, Covid, people, time, job, employee, office, remote, hiring, wellbeing, work life balance and mental health. For all the insights we stated here, employers might need to take care of the whole spectrum of people’s needs and need to look into the working hour regulations during work from home. Deeper analysis of those tweets by gender, age, etc gives us different inputs to the decisions.

Other than shortcomings of work from home we can identify topics such as ‘Work from home recommendations’ and ‘Virtual working’. These topics are with words discovered in line with the Wordcloud such as webinar and virtual assistant. These topics revealed encouragement’s, digital tools, equipment’s, and tips for people to manage both material and immaterial needs will improve support for people’s mental wellness.





The results from the sentimental analyses on overall positivity related to topic also reveals that people are still optimistic about the new normal wfh. Employers and policy makers present their opportunities to ensure that they can continue and rise the satisfaction from their workforce, and they can address critical issues like work and life balance with respect to time.

Finally, other concerns include lack of adequate network , connectivity, infrastructure, improper work environment and drops in people’s physical fitness and lifestyle. Policymakers and employers should consider investing in psychological support for burnt-out employees and improve WFH culture in order to retain their talents and make WFH more efficient.

Overall, positive sentiments are high for all the topics. For “New normal”, the positive sentiment is the highest compared to other topics. The negative sentiment is highest for “Virtual working”. “Virtual working”, “Companies & Culture” (50%) and “Family” (50%) have the lowest proportion of positive tweets and high negative sentiments.