Location Prediction on Twitter Using Machine Learning Techniques

## MAJOR PROJECT PHASE 1 REPORT

***Submitted by***

# Abbas Hussain Muzammil Soma Ram Vighnesh S.Yashwanth Yadav

***in partial fulfillment for the award of the degree of***

## BACHELOR OF TECHNOLOGY

***in***

## COMPUTER SCIENCE AND ENGINEERING



**AURORA’S TECHNOLOGICAL AND RESEARCH INSTITUTE**

## (Affiliated to JNTUH and approved by AICTE, New Delhi) PARVATHAPUR, UPPAL, Hyderabad - 500098

**2021 – 2022**

## AURORA’S TECHNOLOGICAL AND RESEARCH INSTITUTE

**(Affiliated to JNTUH and approved by AICTE, New Delhi) PARVATHAPUR, UPPAL, Hyderabad - 500098**



## DECLARATION

We here by declare that the work described in this project, entitled “**Location Prediction On Twitter Using Machine Learning Techniques”** which is being submitted by us in partial fulfillment for the award of **Bachelor Of Technology in Computer Science and Engineering** to **Aurora’s Technological and Research Institute** is the result of investigation carried out by us under the guidance of **Ms. K. Kavitha, Head of the department, IT.**

The work is original and has not been submitted for any degree of this or any other university.

Place: Hyderabad Date:

**Abbas Hussain Muzammil (18841A0562) Soma Ram Vighnesh (18841A05A8) S.Yashwanth Yadav (19845A0507)**

## AURORA’S TECHNOLOGICAL AND RESEARCH INSTITUTE

**(Affiliated to JNTUH and approved by AICTE, New Delhi) PARVATHAPUR, UPPAL, Hyderabad - 500098**

## BONAFIDE CERTIFICATE

Certified that this project report **“Location Prediction On Twitter using machine learning techniques**” is the bonafide work of “ **Abbas Hussain Muzammil, Soma Ram Vighnesh, S,Yashwant Yadav** who carried out the project under the supervision.

### PROJECT GUIDE MAJOR PROJECT COORDINATOR

**Ms. K. Kavitha Ms. V. Shilpa**

Head of the department Associate Professor

Dept. of IT Dept. of CSE

### HEAD OF THE DEPARTMENT DIRECTOR

**Ms. A Durga Pavani Mr. Srikanth Jatla**

**EXTERNAL EXAMINER**

# Acknowledgement

This work has been done during the project period and it was a very good opportunityto put theoretical knowledge into planned exercise with an aim to solve a real time problem and also to develop confidence to face various practical situations.

We would like to express our gratitude to Director **Mr. Srikanth Jatla**, Aurora’s Technological and Research Institute for providing us with a congenial atmosphere and encouragement.

We express our sincere thanks to Head of the Department **Ms. A Durga Pavani** for giving us the support and her kind attention and valuable guidance to us throughout the course.

We express our sincere thanks to Project Coordinator **Ms. V. Shilpa** for helping us to complete our project work by giving valuable suggestions.

We convey thanks to our project Guide **Ms. K. Kavitha,** Department of Information Technology, for providing encouragement, constant support and guidance which was of great help to complete this project successfully.

# Abstract

Location prediction of users from online social media brings considerable research these days. Automatic recognition of location related with or referenced in records has been investigated for decades. As a standout amongst the online social network organization, Twitter has pulled in an extensive number of users who send a millions of tweets on regular schedule. Because of the worldwide inclusion of its users and continuous tweets, location prediction on Twitter has increased noteworthy consideration in these days. Tweets, the short and noisy and rich natured texts bring many challenges in research area for researchers. In proposed framework, a general picture of location prediction using tweets is studied. In particular, tweet location is predicted from tweet contents. By outlining tweet content and contexts, it is fundamentally featured that how the issues rely upon these text inputs. In this work, we predict the location of user from the tweet text exploiting machine learning techniques namely naive bayes, Support Vector Machine and Decision Tree.

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGENO.** |
|  | **Acknowledgement Abstract**  **List of Figures** | **IV V**  **VIII** |
| **1** | **Introduction** |  |
|  | 1.1 Impact of social media | 1 |
|  | 1.2 Tweet Classification | 1 |
| * 1. Analyzing Tweets   2. Importance of Predicting | 2  3 |
| Location |  |
|  | 1.5 Applications of Location Prediction from Twitter | 3 |
| **2** | **Literature Survey** |  |
|  | 2.1 Existing System | 7 |
|  | 2.2 Proposed System | 9 |
| 2.3 System Study | 11 |
| **3** | **Requirement Specification** |  |
|  | 3.1 Software Requirements | 12 |
|  | 3.2 Libraries Used | 12 |
|  | 3.3 Frameworks Used | 13 |
|  | 3.4 Packages Used | 14 |
|  | 3.5 Hardware Requirements | 15 |
| **4** | **Modules** | 18 |
| **5** | **References** | 22 |

# LIST OF FIGURES

# FIGURE NO. FIGURE NAME PAGE NO.

# 3.4.2 Geopy Servies 15

# 4.8 Decision Tree Model 21

**Chapter 1**

## INTRODUCTION

Twitter is an American micro blogging and social networking service on which users post and interact with messages known as **tweets**. Registered users can post, like, and retweet tweets, however, unregistered users have the ability to only read tweets that are publicly available. Users interact with Twitter through browser or mobile frontend software, or programmatically via its APIs.

Users may post explicitly their location on the tweet text they post, whereas in certain cases the location may be available implicitly by including certain relevant criteria. Tweets are not a strongly typed language, in which users may post casual with emotion images. Abbreviated form of text, misspellings, and extra characters of emotional words makes tweet texts noisy

### Impact of social media

Social media platforms can be efficiently used for supply chain management by professionals, organizations, and retailers for their operations. Social networks like Twitter and Facebook allow users to update information on social activities that they undertake.

### Tweet Classification

The major work in this project is a tweet classification system to classify tweets into high and low priority.

**High Priority tweets** are those, which ask for help, such as food, shelter, medicine etc. during a disaster.

**Example:** Two sample tweets of high priority. Tweet is in the English script, but the words used here are in the Hindi language. The translation of the tweet is, “Mr. @Narendramodi, heavy floods in Chapra Bihar, please arrange for administrative

help, people here are very worried.”

**Low priority tweets** convey information related to a disaster, such as “Rescue team has done a good job.”

**Example** A user thanks Twitter for its help during a disaster.

**The other contribution of this project work is location prediction of high priority tweets**.

### If Geo-tagging information is missing in a tweet

* We predict location by making use of historical Geo-tagged tweets of the specificusers and build a Markov chain.
* The low priority tweets are analyzed to find the spread of the disaster. These mayalso be used to evaluate the performance of different agencies during a disaster.

### Analyzing Tweets

The techniques applied for normal documents are not suited for analyzing tweets. The character limitations of tweets about 140 characters may make the tweet uneasy to understand, if the tweet context is not studied. The issue of location prediction related named as Geo location prediction is examined for Wikipedia and web page documents. Entity recognition from these formal documents has been researched for years. Different types of content and context handling on these documents are also studied extensively. However, the location prediction problem from twitter depends highly on tweet content. Users living in specific regions, locations may examine neighborhood tourist spots, landmarks and buildings and related events.

* + - **Home Location:** User’s residential address given by user or location given by user on account creation is considered as home location. Home location prediction can be used in various application namely recommendation systems, location-based advertisements, health monitoring, and polling etc. Home location can be specified as
      * administrative location
      * geo graphical location or co-ordinates.
    - **Tweet Location:** Tweet location refers to the region from where the tweet is posted by user. By construing tweet location, one can get tweet person’s mobility. Usually, home location is collected from user profile, whereas tweet location can be arrived from user’s Geo tag.

### Importance of predicting location

Social networks, like Twitter and Facebook, are valuable sources to monitor real- time events, such as earthquakes and epidemics. For this type of surveillance, the user’s location is an essential piece of information, but a substantial number of users choose not to disclose their geographical information.

However, characteristics of the users’ behavior, such as the friends they associate with and the types of messages published can hint on their spatial location.

### Applications of Location Prediction from Twitter

When composing tweets, user may make reference to the names of a few locations in tweet texts. Referenced location prediction may encourage better understanding of tweet content, and applications like

* + - recommendation systems
    - location based advertisements
    - health monitoring and polling etc.

The use of social media is being explored as a tool for disaster management by

developers, researchers, government agencies and businesses. The disaster-affected area requires both, cautionary and disciplinary measures For this we need a **computerized decision-making system during emergencies.**

Twitter plays a major role in informing people, acquiring their status of information, and also gathering information on different rescue activities takingplace during both, natural disasters (tsunamis/floods) and man-made disasters (terrorist attack/food contamination)

* + 1. **Why Twitter?**

Twitter provides the space where both official and common people can post their experiences and advice regarding disasters., which makes it a popular choice for disaster management.

* + - * A lot of research work is going on to make this platform more suitable for disaster management.
      * However, a more systematic study of social media is needed to improve public response.

We can appeal to the research community to devise methods to improve citizen- engagement during emergencies. Quick and accurate responses from the leaders during disaster may boost their personal political standing and Several agencies such as **BMKG in Indonesia** are actively engaged in providing updates and warnings to public through Twitter.

Social media is also used by various agencies to coordinate rescue efforts and help victims.

* + 1. **Impact of Twitter**

Twitter is a micro blog where users send brief text messages, photographs and audio clips. Since users write small messages, they regularly send it and check for updates from others.

Twitter updates include social events such as

* + - * Parties
      * Cricket matches
      * Political campaigns
      * Disastrous events such as storms, heavy rainfall, earthquakes
      * Traffic jams etc.

A lot of work has been done to detect events, both social as well as disastrous from Twitter messages. Most disastrous event detection systems are confined to detect whether a tweet is related to the disaster or not, based on textual content. The related tweets are further used to warn and inform people about precautionary measures These tweets are also used to study the tweeting behavior of users during disasters.

We view Twitter not only as an awareness platform, but a place where people can ask for help during disaster.

The tweets asking for help need to be separated from other tweets related to the disaster. These tweets then can be used to guide the rescue personnel.

To help victims in need, one needs to have his/her exact location in their tweet, which is another important issue in emergency situations.

Distribution centers play a big role in helping victims. We can propose a **multi- objective location routing-model** to minimize the cost of opening a distribution center for relief routing.

The real time location estimation plays a big role in logistics, stockpiles, and medical supply planning. The growing number of location-based Social Networks provide the spatiotemporal data that has substantial potential to increase situational awareness and enhance, both planning and investigation .

* + 1. **Twitter Location Prediction Analysis**

The analysis shows that only 26% users mention their location at a city level or below, and the remaining are mostly a country name, or even words with not much meaning, such as Wonderland.

Only 0.42% tweets have geo-tags, but it is found that about 3.17% tweets are geo-

tagged.

These analyses reveal that Twitter has limited applicability as a location-based sensing system.

* + 1. **Utilizing GPS**

The rise of mobile Internet users in the last couple of years has significantly increased the number of mobile twitter users.

According to a report by statista (2020), the mobile Internet users in India will be 749 million by the end of 2020. The same report also highlights the fact that in rural areas, 39% of users are using social media, whereas in urban areas, this percentage is much higher. Mobile Twitter users can switch on and off their geo-tagging, as and when preferred.

The battery power of smart phones plays a significant role here, as the **global positioning system (GPS)** consumes significant amount of batterypower.

Users prefer switching off their GPS to save power.

On the other hand, applications such as taxi hiring services and e-commerce sites such as flipkart.com require GPS to work properly. The analysis of mobile Twitter users thus shows some tweets with geo-tagging, and others without geo-tagging.

* + 1. **Down Fall of Geo Tag**

During emergencies, people want to preserve the battery power of their phones. Hence, tweets with geo-tags will be very few on such occasions.

* + 1. **Challenges in a Country**

For Example: India is a multilingual country, where English is used as the main language for communicating on social media websites. However, users of these sites also use their regional languages.

Hence, event detection in the Indian context also needs to identify variations in the language used.

# Chapter 2

## LITERATURE SURVEY

* 1. **Existing System**
     + In the Existing system to the problem of finding location from social media content. The Social Networks from and motivated by Term frequency (TF) and inverse document frequency (IDF), they arrived Inverse City Frequency (ICF) and Inverse Location Frequency (ILF) respectively.
     + They raked the features by using these frequency values and TF then by TF values. From this they arrived those local words spread in document in few places and have high ICF and ILF values.
     + They approached model for identifying local words indicative or used in certain locations only. They aimed to identify automatically by ranking the local words by their location, and they find their degree of association of location words associated to particular location or cities.
     + As a standout amongst the online social network organization, Twitter has pulled in an extensive number of users who send a millions of tweets on regular schedule. Because of the worldwide inclusion of its users and continuous tweets, location prediction on Twitter has increased noteworthyconsideration in these days. In this work, we predict the location of user from the tweet text exploiting machine learning techniques namely naive bayes, Support Vector Machine and Decision Tree [1].
     + Social networks, like Twitter and Facebook, are valuable sources to monitor Real - time events, such as earthquakes and epidemics. For this type of surveillance, the user’s location is an essential piece of information, but a substantial number of users choose not to disclose their geographical information. However, characteristics of the users’ behavior, such as the friends they associate with and the types of messages published can hint on their spatial location. In this, we present a method to infer the spatial location of Twitter

users. Unlike the approaches presented so far, we incorporate two sources of information to learn the geographical position: the text posted by users and their friendship network. We propose a probabilistic approach that jointly models the geographical labels and the Twitter texts of the users organized in the form of a graph representing the friendship network. We use Markov random field probability model to represent the network and learning is carried out through a Markov chain Monte Carlo simulation technique to approximate the posterior probability distribution of the missing geographical labels. We demonstrate the utility of this model in a large dataset of Twitter users, where the ground truth is the location given by GPS dispositive. The method is evaluated and compared to two baseline algorithms that user these two types of information separately, the accuracy rates obtained are significantly better than those of the baseline methods. Index Terms—Network Learning; Geographic Targeting; Geolocation Estimation; Spatial Data Mining [2].

* + - Getting data comes as the second step in any data science/machine learning project lifecycle, right after framing the problem you want to solve, which would make this step be the backbone of the rest of the phases. Also, socialmedia are great places to collect data, especially for competitor analysis, topic research, sentiment analysis, etc. This article aims to perform a step-by-step implementation on how to get credentials and the implementation on a simple use case [3].
    - Geo location prediction is vital to geospatial applications like localized search and local event detection. Predominately, social media geo location models are based on full text data, including common words with no geospatial dimension (e.g. today) and noisy strings (tomorrow), potentially hampering prediction and leading to slower/more memory-intensive models. In this paper, we focus on finding location indicative words (LIWs) via feature selection, and establishing whether the reduced feature set boosts geo location accuracy. Our results show that an information gain ratio based approach surpasses other methods at LIW selection, outperforming state-of-the-art geo location prediction methods by 10.6% in accuracy and reducing the mean and median of prediction error

distance by 45km and 209km, respectively, on a public dataset. We further formulate notions of prediction confidence, and demonstrate that performance is even higher in cases where our model is more confident, striking a trade-off between accuracy and coverage. Finally, the identified LIWs reveal regional language differences, which could be potentially useful for lexicographers [4].

* + - From the book we can Understand the importance of applying spatial relationships in data science, Select and apply data layering of both raster and vector graphics, Apply location data to leverage spatial analytics, Design informative and accurate maps, automate geographic data with Python scripts, Explore Python packages for additional functionality, Work with atypical data types such as polygons, shape files, and projection, Understand the graphical syntax of spatial data science to stimulate curiosity [5].
    - You are Here is a compelling read perfect for anyone interested in learning about the history of location technology. This book examines how having constant access to location data has come to be a part of modern life, and highlights the key players in the development of location and mapping technologies. Informative and interesting, you are Here discusses how solving one problem can sometimes usher in a new set of concerns [6].
  1. **Proposed System**
     + Live stream of twitter data is collected as dataset using authentication keys.

The aim of proposed system is to predict the user location from twitter content considering

**–** user home location,

-- tweet location

-- tweet content.

* + - To handle this, we used three machine learning approaches to make prediction

easier and finding the best model amongst them. Live tweet stream from twitter for keyword “apple” is collected and stored in Tweet table.

* + - live twitter data can be collected by registering a consumer key, consumer

\_secret, access token, access token secret for authentication and collecting live stream of tweets. we have collected more than 1000 tweets of particular keywords such as Indian city hashtag names. We can also search tweets based on hashtag.

## System Study

* + 1. **Feasibility Study**

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

* + - * **Economic Feasibility:** This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.
      * **Technical Feasibility:** This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system de-- veloped must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.
      * **Social Feasibility:** The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must

be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

# Chapter 3

## REQUIREMENT SPECIFICATION

### Software Requirements

Following are the Software’s required followed by their detailed description

### Python

* + - * Python is a high-level, general-purpose and a very popular programming language.
      * Python programming language (latest Python 3) is being used in web develop- ment, Machine Learning applications, along with all cutting-edge technology in Software Industry.
      * Python allows programming in Object-Oriented and Procedural paradigms.
      * Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.
      * Python is dynamically typed and has automatically garbage is collected.
      * It supports multiple programming paradigms, including procedural, object- oriented, and functional programming

.

### Operating System

An operating system (OS) is system software that manages computer hardware, software resources, and provides common services for computer programs.The operating system should support Python 3.7.0 and all the libraries mentioned above.

### Libraries Used

* + 1. **Pandas**

Pandas is an open-source library that is used for data analysis and manipulation. It is

a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance productivity for users. In this project we have used pandas version 0.25.3.

### Numpy

Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. In this project we have used Numpy version 1.81.1.

### Sklearn

Sklearn is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

### Frameworks Used

* + 1. **Django**

The Django web framework is a free, open source framework that can speed up development of a web application being built in the Python programming language.

Django is designed in such a way that encourages developers to develop websites fast, clean and with practical design. Django’s practical approach to getting things done is where it stands out from the crowd.

If we are planning to build a highly customize app, such as social media website, Django is one of the best frameworks to consider. Django strength lies in its interaction between users or its ability to share different types of media. One of the great advantage of django is its ability to utilize large community-based support which gives us a highlycustomize third-party ready to use plugins in your applications.

### Why Django?

* + - * It allows developers to use modules for faster development.
      * It has built-in protection for some common security issues such as cross-site scripting, request forgery, click jacking, and SQL injection. Django releases

new security patches quite often and it immediately responds to the security vulnerabilities and alerts other frameworks.

### Packages

* + 1. **Tweepy**

Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python.

It gives you an interface to access the API from your Python application

### Geopy

Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

The geopy package is not geocoding service provider. It just provides an interface to connect to several services under a single package.

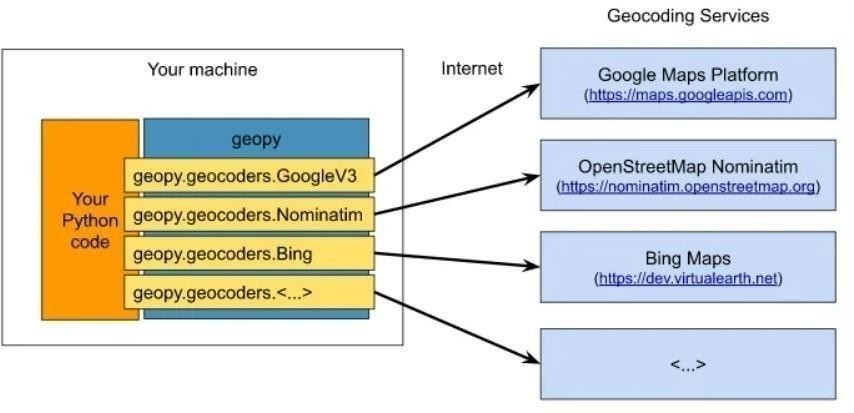


Figure 3.4.2: Geopy Services

### Hardware Requirements

For developing the application, following are the Hardwar Requirements

* **Processor**: Intel i3 or above, AMD Ryzen 3 or above
* **RAM:** 8 GB Minimum
* **Hard Disk:** 5 GB free hard disk space

# Chapter 4 Modules

### Data Collection

Live tweet stream from twitter for keyword “apple” is collected and stored in 'twitter.json’ file. Live twitter data can be collected by registering a consumer\_key, consumer\_secret, access\_token, access\_token\_secret for authentication and collecting live stream of tweets. We have collected more than 1000 tweets of particular keyword such as ‘Chennai, Mumbai and Kerala’. The information extracted from live includes tweetid, name, screen\_name, tweet\_text, HomeLocation, TweetLocation, MentionedLocation, Lvalue.

### Data Preprocessing

* + - Extra characters are removed from tweet text.
    - Capitalize all words to find for geo location .
    - Remove the tweet if user home location not mentioned.
    - Mention home location in tweet location, if user tweet location is null.
    - Removes tweets if no location is mentioned in tweet text.
    - Final extract geodata from tweet text.

### Search Tweets

In this module we can enter HASHTAG and then application will search tweets from twitter by using TWEEPY API to read all tweets from given tweet.

### Geopy

Geopy makes it easy for Python developers to locate the coordinates of addresses

,cities, countries, and landmarks across the globe using third-party geocoders and other

data sources. The geopy package is not geocoding service provider. It just provides an interface to connect to several services under a single package.

### Tweepy

Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python. It gives you an interface to access the API from your Python application.

### Naive Bayes Classification

Naive Bayes classifier is the most popular and simple classifier model used commonly. This model finds the posterior probability based on word distribution in the document. Naïve Bayes classifier work with Bag Of Words(BOW) feature extraction model, which do not consider the position of word inside the document. This model used Bayes Theorem for prediction of particular label from the given feature set. The dataset is split into train set and test set. Upon test set, NB\_model is applied to find the location prediction.

There are three types of Naive Bayes Model, which are given below:

* **Gaussian**: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
* **Multinomial**: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education etc. The classifier uses the frequency of words for the predictors.
* **Bernoulli**: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks

### Support vector machine

Support vector machine is one of most common used supervised learning techniques, which is commonly used for both classification and regression problems. The algorithm works in such a way that each data is plotted as point in n dimensional space with the feature values represents the values of each co-ordinate.

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

The followings are important concepts in SVM –

* **Support Vectors** − Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.
* **Hyperplane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
* **Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps –

* First, SVM will generate hyperplanes iteratively that segregates the classes in best way.
* Then, it will choose the hyperplane that separates the classes correctly

### Decision Tree

### Decision tree is the learning model, which utilizes classifications problem. Decision tree module works by splitting the dataset into minimum of two sets. Decision tree’s internal nodes indicates a test on the features, branch depicts the result and leaf ’s are decisions made after succeeding process on training.

Decision Tree works as follows

* Decision tree starts with all training instances linked with the root node.
* It splits the dataset into train set and test set.
* It uses information to gain and chooses attributes to label the each node. Subsets made contain information with a similar feature attribute.
* Above process is repeated till in all subset until leafs get generated in tree.

Attribute Selection Measures –

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.**By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are

* Information Gain

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first.

* Gini Index

Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.

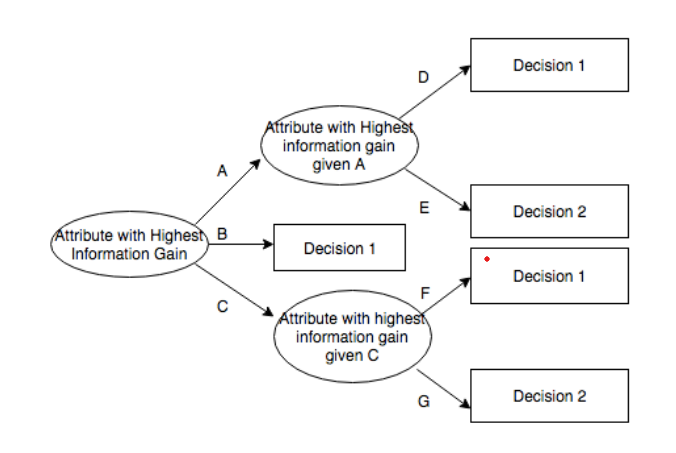


Figure 4.8: Decision Tree Model

# Chapter 5

# References

1. “Location Prediction on Twitter Using Machine Learning” By Indira K.Brumancia and E. Siva Kumar DOI: 10.1109*/*ICOEI.2019.8862768.
2. “Uncovering the location of Twitter User ” By Renato Assunc and Wagner Me- ira Jr. , DOI: 10.1109*/*BRACIS.2013.47.
3. “Collect Data From Twitter : A step by step implementation using tweepy” By Zoumana Keitahttps ,:*//*towardsdatascience.com*/*collect-data-from-twitter-a-step-by-step- implementation-using-tweepy-7526*ff*f2cb31.
4. “Geo location Prediction in Social Media Data by Finding Location Indicative Words” By Han, Cook & Paul Baldwin [https://aclanthology.org/C12-1064.](https://aclanthology.org/C12-1064)
5. “Python for Geospatial Data Analysis” By Bonny P. McClain.

### “You are Here: From the Compass to GPS, the History and Future of How We Find Ourselves” By Hiawatha Bray.