

**PREDICTING ONLINE EXTREME RELIGIOUS DISCOURSE USING
NATURAL LANGUAGE PROCESSING TECHNIQUE**

BY

**MD. SALMAN KAISER
ID:171-15-1207
AND**

**JIHAD AZAD JISAN
ID: 171-15-1209
AND**

**MONTASIR MAHFUZ MAHIN
ID:171-15-1344**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sabab Zulfiker
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Saif Mahmud Paevez
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2021

APPROVAL

This Project titled “Predicting Online Extreme Religious Discourse Using Natural Language Processing Technique”, submitted by Md. Salman Kaiser, Jihad Azad Jisan and Montasir Mahfuz Mahin to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *14-1-2021*.

BOARD OF EXAMINERS

Professor Dr. Touhid Bhuiyan

Professor and Head

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Chairman

Dr. S. M. Aminul Haque

Associate Professor and Associate Head

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Mr. Ohidujjaman

Senior Lecturer

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Technology

Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sabab Zulfiker, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Md. Sabab Zulfiker
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Saif Mahmud Parvez
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Md. Salman Kaiser
ID: 171-15-1207
Department of CSE
Daffodil International University

Montasir Mahfuz Mahin
ID: 171-15-1344
Department of CSE
Daffodil International University

Jihad Azad Jisan
ID: 171-15-1209
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I thank and wish my supervisor's profound awareness and a keen interest in carrying out this project in the area of "*Natural Language Processing*." We now consider it possible to complete this project with her infinite patience, academic guidance, relentless motivation, energetic supervision, constructive criticism, precious advice, and correcting them at all times.

We would like to express our heartiest gratitude to **Md. Sabab Zulfiker, Lecturer**, Department of CSE Daffodil International University and **Dr. S.M Aminul Haque, Associate Professor & Associate head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In the current age innovation has become an indistinguishable piece of our life. Interpersonal interaction Site is an incredible advancement of current occasions. Facebook, Twitter and so forth have become an ordinary piece of people groups' lives. Everyone utilizing the web these days utilizes long range informal communication destinations. Online media has become a stage for each sort of correspondence. Presently-a-days one can scarcely discover any individual who isn't a client of any online media. Web-based media calculations, today, work in a way where one as a rule sees the sort of posts one prefers or is lined up with, making the scope of discussions smaller and, frequently, and their unnecessary utilization risky. Just as the different employments of interpersonal interaction locales, individuals in some cases end up engaged with genuine viciousness, incited by some online media posts or exercises. One of them is strict viciousness. Strict viciousness is a term that covers marvels where religion is either the subject or the object of rough conduct. Strict viciousness is brutality that is inspired by, or in response to, strict statutes, messages, or the principles of an objective or an aggressor. It incorporates viciousness against strict foundations, individuals, items, or occasions. Strict savagery doesn't solely allude to acts which are submitted by strict gatherings, all things being equal, incorporates acts which are submitted against strict gatherings. Strict savagery is going through a restoration. The spike in exacting mercilessness is worldwide and impacts fundamentally every severe social affair. Nowadays individuals are utilizing interpersonal interaction destinations for posting or remarking their talks about religion which have positive or pessimistic effect and may be answerable for religion brutality. The study focused on religion discourses from some popular social media sites and would have predicted extreme religious discourse data among them.

TABLE OF CONTENTS

CONTENTS	PAGE
Acknowledgement	i
Abstract	ii
List of Figures	iii
List of Tables	iv
CHAPTER 1: Introduction	11-17
1.1 Introduction	11
1.2 Motivation	15
1.3 Rational of The Examination	15
1.4 Research Question	16
1.5 Research Organization	17
CHAPTER 2: BACKGROUND	18-21
2.1 Related Work	18
2.2 Scope of The Problems	19
2.3 Challenges	20

CHAPTER 3: RESEARCH METHODOLOGY	22-44
3.1 Introduction	22
3.2 Algorithm Description	22
3.3 Proposed Model	40
3.4 Data Collection Procedure	41
3.5 Implementation and Requirements	44
 CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	 45-51
4.1 Introduction	45
4.2 Experiment Process	45
4.2 Experimental Result	50
4.3 Summary	51
 CHAPTER 5: CONCLUSION AND FUTURE SCOPE	 52
5.1 Conclusion	52
5.2 Future Scope	52
 REFERENCES	 53-55

LIST OF FIGURES

FIGURES	PAGE NO
Fig 1.1.1: Religious issues conflict rate	12
Fig 1.1.2: Communal Peace	13
Fig 1.1.3: Sentiment Analysis	14
Fig 3.2.1: Bayes Theorem Example	23
Fig 3.2.2: Bayes rule	24
Fig 3.2.3: Normal Distribution	25
Fig 3.2.4: Gaussian rule	26
Fig 3.2.5: Normal Dataset	27
Fig 3.2.6: Random forest decision tree	28
Fig 3.2.7: Random Forest Tree	29
Fig 3.2.8: Real life scenario	31
Fig3.3.9: How decision tree works	32
Fig3.3.10: Impurity	34
Fig 3.2.11: Avg Impurity Rule	35
Fig 3.2.12: Example	36
Fig:3.2.13: SVM entering	37
Fig 3.2.14: Sample Data Set	38
Fig 3.2.15: Sample Data Set	39
Fig 3.3.1: Working Process flowchart	41
Fig 3.4.1: Collecting data from social media	42
Fig 3.4.2: Data collection and store	42
Fig:3.4.3: Percentage of positive and negative data	43
	46

Fig 4.2.1: Value of Naïve Bayes Classifier	
Fig 4.2.2: Value of Random Forest Classifier	47
Fig 4.2.3: Value of Decision tree Classifier	48
Fig 4.2.4: Value of SVM classifier	49
Fig 4.3.1: Comparison of different classifier	51

LIST OF TABLES

TABLES	PAGE NO
Table3.2.1: Dataset	32
Table 3.4.1: Data cleaning process	44
Table 4.3.1: The results in different classifier	50

CHAPTER 1

Introduction

1.1 Introduction

Natural language processing technique enables computers to speak with humans of their own language and scales other language-associated duties. NLP makes it workable for PCs to understand text, to hear discourse, decipher it, measure sentiment assumption and figure out which parts are significant. Our study is based on predicted extreme religious discourse using NLP techniques. The improvement of web-based media to offer space to the client or network of articulation, sentiment and articulation unreservedly. Netizens can easily elicit their freedom of expression through web based societal sites. These sites are now most popular for expressing people's opinions. Feeling investigation is chiefly worried about the ID and grouping of conclusions or feelings of each comment on social sites. Slant examination is extensively characterized in the two kinds. The initial one is an element or viewpoint-based notion investigation and the other is objectivity-based notion examination. The predicted extreme religious discourse is the classification of the component-based notion examination. Objectivity based supposition examination does the investigation of the comments which are identified with the feelings like disdain, miss, love and so on. In our study we will reduce the extreme religious discourse using Natural Language Processing technique.

Religious extremism is as of now a fervently discussed point, it is frequently diminished to a unidimensional build that is connected to strict savagery. We contend that the contemporary utilization of the expression "extraordinary" neglects to catch the various translations, convictions, and mentalities characterizing outrageous strict personality. In this issue, we reveal the significance of the expression "extraordinary" in strict settings and answer the call by researchers to give a more thorough structure that fuses the various

measurements that comprise religion. We have built a framework in which we can predict extreme religious discourse using NLP technique. Sometimes in web based social sites or blogging sites people post some inappropriate content that hurts others religious sentiment and create violence and, on that post, then the comments are blown out of proportion and create a mess. Then it affects our internet world and our real life very badly. People became very furious or violent for this kind of comments and posts on social media. It affects our peaceful society very badly. These kinds of comments on social media spoil our communal harmony and create riots.

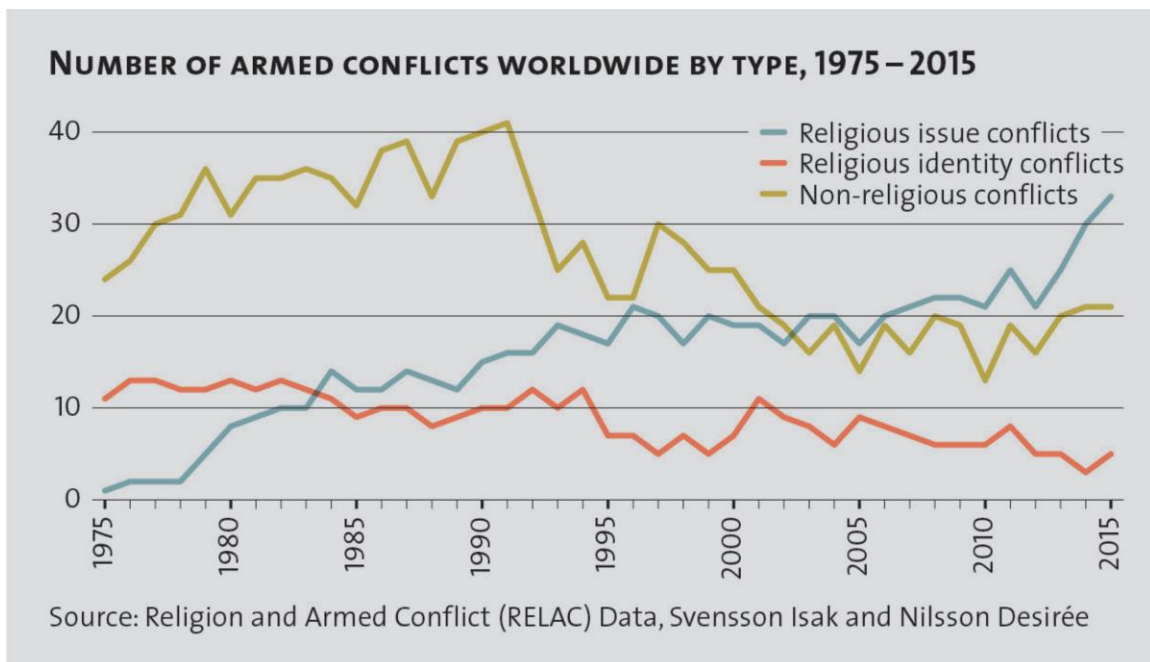


Fig 1.1.1: Religious issues conflict rate

So that we have to find out these extreme religious discourse comments so we build some solutions using NLP techniques and find out solutions. Through our project we can find out inappropriate comments and try to remove those comments from social media for our communal peace.



Fig 1.1.2: Communal Peace

Purposes of our study was to play out a conclusion investigation of information about Uber sourced from Facebook. The reason for the conclusion investigation covering the time was to uncover changes in the manner how extreme religious discourse is seen by. Religion is one of the most sensitive aspects towards the people. Facebook should be taken a gander at as a tremendous information base, which contains significant suppositions and perspectives on its users. There are too many techniques in NLP. We use a few techniques they are Naive bayes, Random Forest, Decision Tree, and SVM to find out extreme religious discourse in societal platforms. By and large, different representative strategies and NLP methods are utilized to break down the conclusion from the societal platform's information. So, in another manner we can say that a supposition examination is a framework or model that takes the reports that dissected the info, and produces an itemized archive summing up the assessments of the given input record. In the initial step pre-preparing is finished. In the first step we are converting the uppercase letter to the lowercase then eliminating the stop words, blank areas, rehashing words, emojis and #hash labels. To effectively characterize the comments, NLP strategy utilizes the preparation information. Thus, this procedure does not need the information base of words like utilized in information-based methodology and subsequently, NLP strategies is better and quicker. Not need the information base of words like utilized in information-based methodology and subsequently, NLP strategies is better and quicker.

Estimation investigation is classified into three various levels which are record level, sentence level and element angle level. In general feeling is to be recognized in record level examinations. Just assessment of sentence is to be recognized in sentence level investigation. Zero in is straightforwardly on feeling itself in element perspective level examination. The few techniques are utilized to remove the component from the source text. Highlight extraction is done in two stages: In the primary stage extraction of information identified with Facebook and YouTube is done for example their particular information is separated. Presently by doing this, the remarks are changed into typical content. In the following stage, more highlights are extricated and added to include vectors. Each tweet in the preparation information is related to the class name. This preparation information is passed to various classifiers and classifiers are prepared. At that point test remarks are given to the model and order is finished with the assistance of these prepared classifiers. So, at long last we get the tweets which are arranged into the positive and negative.



Fig 1.1.3: Sentiment Analysis

1.2 Motivation: Conclusion investigation to recognize religion savagery is a gainful subject of examination. Everyone utilizing the web these days utilizes long range informal communication locales. Individuals share their suppositions and assumptions on the web each day. Which are genuine and not made up. Conclusions are a typically abstract articulation that portrays an individual's supposition, emotions towards the article or administration.

Individuals once in a while end up associated with genuine viciousness, incited by some online media posts or exercises. Religion viciousness is one of them. There are so numerous well-known long-range informal communication locales of present time. Clients can undoubtedly communicate their contemplations as post and remark. Some of them have positive or negative effects and may be liable for religion riots. There are some extreme occurrences that have occurred for spreading strict brutality in online destinations. By our undertaking we can examine that information and foresee extraordinary strict talks among them. So far assessment investigation is performed just in business, political areas and so on No exploration has been done at this point to identify extraordinary strict discourse against a religion or a strict gathering. Along these lines, we intend to take a shot at this area.

1.3 Rational of the examination: In the period of current science and innovation individuals are investing a large portion of their energy in web and long-range informal communication destinations. Individuals utilize interpersonal interaction locales for convey and express their ideas as post and remarks. Some of the time individuals post or offer their talks about religion that harms others' strict assessment and make revolt. We have utilized some Natural Language Processing (NLP) procedures, for example, Naïve Bayes, Random Forest, Decision Tree and SVM to discover extraordinary strict talks from strict remarks and posts of web-based media.

Our examination targets are given underneath: -

- arrange whether a discourse contains contempt towards a religion gathering.
- distinguish whether the discourse spreads brutality among various religion gatherings.
- decide the words that are usually utilized for speeding extraordinary strict talk.
- look at the grouping aftereffect of various classifiers utilizing the dataset to locate the best one.

1.4 Research Question:

In our research several questions can be asked. Such as:

- 1.How can classify extreme religious discourse from social media?
- 2.How does this research help the society?
- 3.What is the future scope of this research?
- 4.Who will benefit from this research?
- 5.Does anyone research on this topic before using different algorithms?
- 6.Which algorithm can be more effective in this research?

1.5 Report Organization: This report contains 5 chapters. They are Introduction, Background, Research and Methodology, Experimental Results and Discussion, Conclusion and Future Scope and at last we discuss the Reference that which articles and sources we read for making this report. In the introduction part we discuss 1.1 intro, 1.2 Motivation, 1.3 Rationale of the study, 1.4 research question and 1.5 research organization.

In the Background chapter we discuss 2.1 Related works, 2.2 Scope of the problem and 2.3 challenges.

In the research methodology chapter, we discuss 3.1 introduction, 3.2 algorithm discussion, 3.3 proposed model, 3.4 data collection procedure, 3.5 Implementation and requirements.

In the Experimental and result chapter we discuss 4.1 Introduction, 4.2 Experimental Result and 4.3 Summary.

In the the conclusion and future scope chapter we discuss 5.1 Summary of the study, 5.2 Conclusion and 5.3 Future Scope.

And at last we gave the references.

CHAPTER 2

BACKGROUND

2.1 RELATED WORK:

We have studied about sentiment analysis on social media data, as very little work has been done in this field. Predicting positive and negative data of social media about religion may reduce religious violence. This field needs a lot of work to be done. For doing this work we have studied some work which is related to predict sentiment.

Taimur Islam et al. [12] have worked in machine learning and prepared a dictionary consisting of unique words collected from political or nonpolitical posts or comments and then trained using Naïve Bayes algorithm based on probability theory. They have tested their algorithm using 200 postings from Facebook and their result shows that the method can classify posts or comments with 83% accuracy.

Akshay Amolik et al. [2] have worked with the help of feature vector and classifiers such as Support vector machine and Naïve Bayes, they were correctly classifying these tweets as positive, negative and neutral to give sentiment of each tweet. They get 75% accuracy from SVM and 65% accuracy from Naïve Bayesian classifier.

Akshma Chadha et al. [3] have applied different machine learning algorithms like Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Decision Tree, Logistic Regression, and Support Vector Machine on the data to predict and identify trends of suicidal ideation. They get accuracy of Multinomial Naïve Bayes 76.67% Bernoulli Naïve Bayes 78.42% Decision Tree 79.30%, Logistic Regression 79.65%, Support Vector Machine 79.30%, Voting Ensemble 79.65%, AdaBoost Ensemble 79.47%, and Random Forest 78.77%.

Efthymios Kouloumpis et al. [7] investigated the utility of linguistic features for detecting the sentiment of Twitter messages. Their work of Data preprocessing consists of three

steps: 1) tokenization, 2) normalization, and 3) part-of-speech (POS) tagging. The number n-grams to include as features was determined empirically using the training data. This is equivalent to 10% of the training data. They experimented with different sample sizes for training the classifier, and this gave the best results based on the validation data. The rounds of boosting was determined empirically using the validation set. They also experimented with SVMs, which gave similar trends, but lower results overall.

Anna Baj-Rogowska [4] used machine learning, lexicon based, hybrid approach. Data used as a basis for the analysis were opinions expressed by Facebook users about Uber and collected in the period between July 2016 and July 2017. The accuracy of their work was 85%.

Jibon Naher et al. [9] gave an overview about the alarming phenomenon of real life violence provoked by Facebook activates in Bangladesh.

R. A. S. C Jayasanka et al. [11] have worked with Naïve Bayes Classifier, Maximum Entropy Classifier. After evaluating the results of the evaluation process, as the most suitable supervised learning method from accuracy wise and performance wise, it selected the Naïve Bayes classifier to classify sentiments with most informative unigrams and bigrams as the feature extraction method.

2.2 Scope of the problem:

Assessment examination is one of the most well-known subjects of exploration everywhere in the world, since till now however a ton of work has been done yet there is an absence of precision also, understanding for the machine. Along these lines' individuals are buckling down on the calculation, conclusion word library and various methods. In this exploration we are utilizing a calculation-based method known as Naïve Bayes calculation. Our center will be the improvement of the precision.

2.3 Challenges:

One of the prime issues of assumption expectation currently is the exactness. A few specialized or then again calculated difficulties become obstructions in investigating the exact significance of conclusions and recognizing the reasonable slant extremity. To recognize and remove abstract data from text the conclusion investigation is the act of applying common language preparing and text examination methods.

The level of precision issue is difficult to reply, said Bing Liu, a University of Chicago software engineering educator represent considerable authority in information mining. It relies upon what are estimating, the degree of investigating text, and the quantity of informational indexes across spaces and the voice sound nature of recordings, among different factors. All things considered; he imagines that progress is being made in such a manner. In estimation expectation it is additionally testing to distinguish a more inside and out estimation/feeling. Positive and negative is an exceptionally basic examination however the difficult one is to separate feelings like how much scorn there is inside the sentiment, how much joy, how much trouble, and so on.

Feeling discovery is actually a troublesome undertaking on the grounds that occasionally it happens that somebody tells something that appears to be positive however in genuine it's not positive the sense was negative. So, some of the time it is hard to comprehend the significance of a sentence because the feelings are an excess of complex.

Mostly conclusion forecasts attempt to identify the psychological circumstance of an individual. Be that as it may some of the time it become intense to determine what the individual implied. In the event that we think about sound opinion investigation, at that point clamor or voice tune distinction can make significant blunders in yield. Same for text examination, since certain writings word astute significance is entirely unexpected from its genuine significance. That is the reason supposition investigation is confronting significant difficulties nowadays.

CHAPTER 3

METHODOLOGY

3.1 Introduction:

In our study we have to predict extreme religious discourse from social media. From several techniques of NLP, we use four most effective algorithms for our work. Now we will elaborately discuss these algorithms. I will do most of my viable work. To begin my work, I need some information to take a shot at. As I have begun as of now, I will utilize social media information as our data as our informational collection. For gathering information, I use social media and various kinds of websites. When the information has been gathered it will be saved in natural crude configuration. To get fairly precise outcome from my program I need to pre-measure the information and make it usable for our work.

3.2 Algorithm Description: In our study we have to predict extreme religious discourse from social media. From several techniques of NLP, we use four most effective algorithms for our work. Now we will elaborately discuss these algorithms.

Naive Bayes: This classifier accepts that the presence of a specific component in a class is irrelevant to the presence of some other element. Regardless of whether this element relies upon one another or upon the presence of different highlights these properties autonomously add to the likelihood whether an organic product is an apple or an orange or a banana so that is the reason it is known as a credulous base. This is a straightforward yet shockingly amazing calculation. It is extremely simple to construct and especially valuable for enormous datasets. The likelihood and insights hypothesis which is then again known as the bayes law or the bayes rule portrays the likelihood of an occasion dependent on earlier information on conditions that may be identified with the occasion bayes hypothesis is an approach to sort out contingent likelihood. The restrictive likelihood of an occasion happening given that it has some relationship to at least one different occasion for instance

your likelihood of getting a parking spot is associated with the time you park where you park and what shows are you going on at that time. Bayes hypothesis is somewhat more nuanced, more or less it gives you a characteristic likelihood of an occasion given data about the test. Presently on the off chance that you take a gander at the meaning of bayes hypothesis we can see that given a speculation H and the proof E bayes hypothesis expresses that the connection between the likelihood of the theory prior to getting the proof which is the P of H and the likelihood of the theory subsequent to getting the proof is P of H given E is characterized as likelihood of E given H into likelihood of S isolated by likelihood of E $P(H|E)=P(E|H).P(H)/P(E)$. It's somewhat befuddling right? So we should take a guide to comprehend this hypothesis so assume

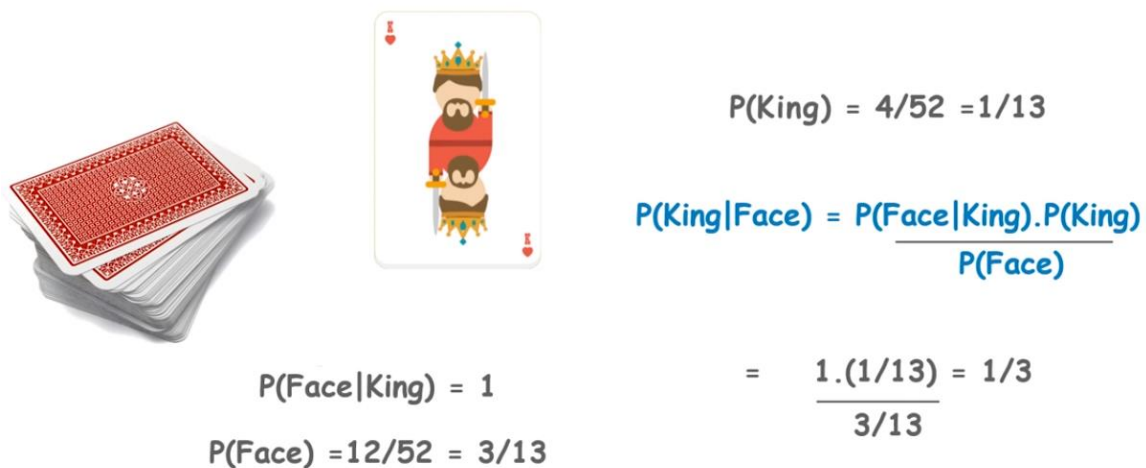


Fig 3.2.1: Bayes Theorem Example

I have a deck of cards and if a solitary card is drawn from the deck of playing a game of cards the likelihood that the card is a lord is 4/52 since there are four rulers in a standard deck of 52 cards now if ruler is an occasion this card is a lord the likelihood of a lord is given as 4/52 that is equivalent to 1/13. Presently if the proof is accommodated example somebody takes a gander at the card that the single card is a face card the likelihood of a ruler that its face can be determined utilizing the bayes hypothesis by this recipe. Presently since each is likewise a face card the likelihood of face given that it's a lord is equivalent

to 1. What's more, there are 3 face cards in Each suit is the jack lord sovereign. The likelihood of the face card is equivalent to 12/52 that is 3/30. Presently utilizing Bayes hypothesis, we can discover the likelihood of the lord that it's a face so our last answer comes which is 1/3. So, this is the basic illustration of a bayes hypothesis.

So, the bayes hypothesis law is given underneath shoes in the fig

Likelihood
How probable is the evidence
Given that our hypothesis is true?

Prior
How probable was our hypothesis
Before observing the evidence?

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Posterior
How probable is our Hypothesis
Given the observed evidence?
(Not directly computable)

Marginal
How probable is the new evidence
Under all possible hypothesis?

Fig 3.2.2: Bayes rule

Sorts of Naive Bayes Classifier:

Multinomial Naive Bayes: This is commonly used for report game plan issue, i.e., whether a record has a spot with the class of sports, administrative issues, advancement, etc. The features/pointers used by the classifier are the repeat of the words present in the record.

Bernoulli Naive Bayes: This resembles the multinomial honest bayes yet the pointers are Boolean elements. The limits that we use to anticipate the class variable take up characteristics yes or no, for example if a word occurs in the substance or not.

Gaussian Naive Bayes: Exactly when the pointers take up a predictable worth and are not discrete, we acknowledge that these characteristics are tried from a gaussian scattering.

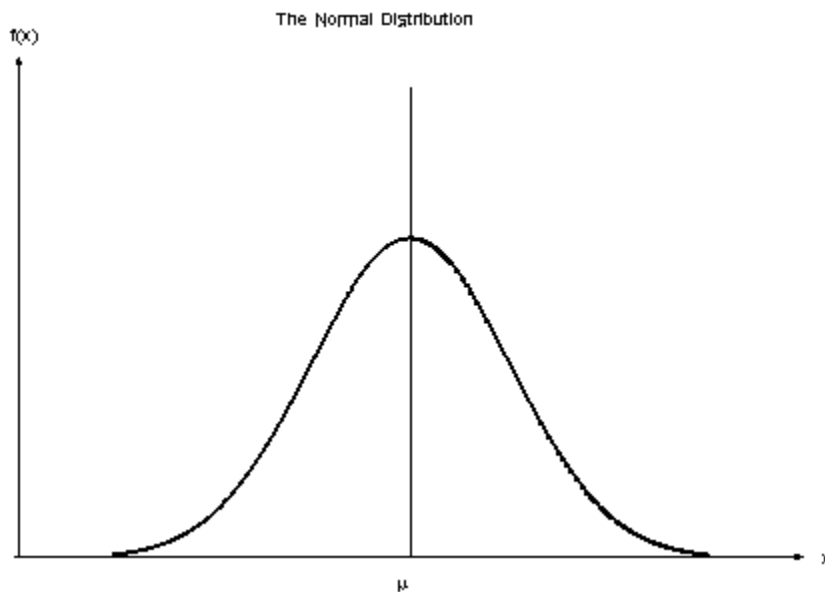


Fig 3.2.3: Normal Distribution

Since the manner in which the qualities are available in the dataset changes, the recipe for contingent likelihood changes to,

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Fig 3.2.4: Gaussian rule

How Naive Bayes calculation functions?

How about we comprehend it utilizing a model. Beneath I have a preparation informational index of climate and relating objective variable 'Play' (proposing potential outcomes of playing). Presently, we need to arrange if players will play dependent on climate conditions. We should follow the underneath steps to perform it.

Stage 1: Convert the informational index into a recurrence table

Stage 2: Create Likelihood table by finding the probabilities like Overcast likelihood = 0.29 and likelihood of playing is 0.64.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Fig 3.2.5: Normal Dataset

Stage 3: Now, utilize Naive Bayesian condition to ascertain the back likelihood for each class. The class with the most elevated back likelihood is the result of expectation.

Issue: Players will play if the climate is bright. Is this assertion, right?

We can comprehend it utilizing the above examined technique for back likelihood.

$$P(\text{Yes} \mid \text{Sunny}) = P(\text{Sunny} \mid \text{Yes}) * P(\text{Yes})/P(\text{Sunny})$$

Here we have $P(\text{Sunny} \mid \text{Yes}) = 3/9 = 0.33$, $P(\text{Sunny}) = 5/14 = 0.36$, $P(\text{Yes}) = 9/14 = 0.64$

Presently, $P(\text{Yes} \mid \text{Sunny}) = 0.33 * 0.64/0.36 = 0.60$, which has higher likelihood.

Guileless Bayes utilizes a comparative technique to anticipate the likelihood of various class dependent on different traits. This calculation is generally utilized in content order and with issues having different classes.

Random Forest: These calculations work by building developing various choice trees during preparing phase. The choice of most of the trees is picked by the irregular timberland as an official choice.

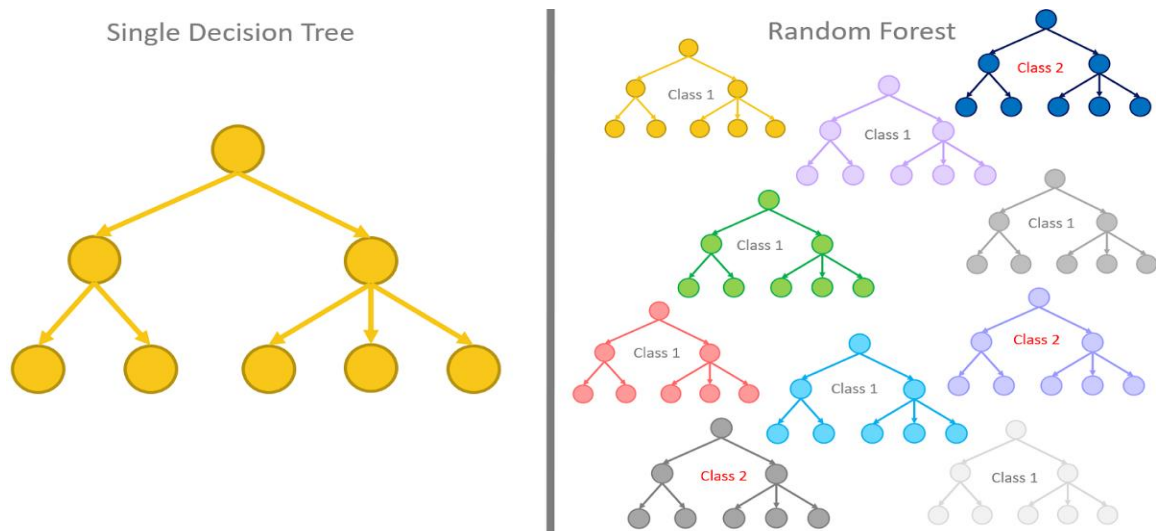


Fig 3.2.6: Random forest decision tree

Why do we utilize irregular backwoods since it has some preferred position to conquer our task? Initial one is it has no overfitting. Overfitting implies we have fit the information so near what we have fit the information so near what we have is our example that we get all the abnormal parts and as opposed to foreseeing by and large information you are anticipating the odd staff which you don't need. its preparation time is less. High exactness round huge information base productively. For huge information base it delivers exceptionally exact predictions. It gauges missing information. In today's world information is chaotic so when you have an arbitrary backwoods it can keep up exactness when enormous extent information is missing what that implies when have information from five and various regions and they took one insights in a single zone and they took a marginally extraordinary arrangement of measurements in the other so they a portion of a similar shared information yet one is missing like the quantity of offspring of the house on the off chance that you accomplishing something over segment and other one is feeling the loss of the size of a house it will look both of those independently and fabricate two distinct

trees and it can do an excellent employment of speculating which one fits better despite the fact that its missing the information.

Example:

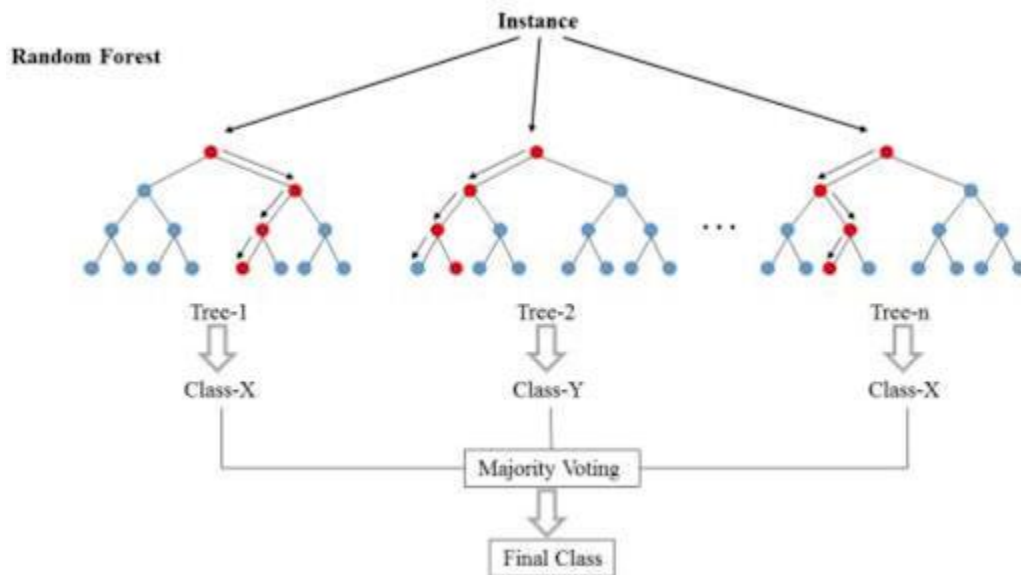


Fig 3.2.7: Random Forest Tree

Assume irregular choice trees anticipate nearly three novel targets, for example, class_X = finger, class-Y= thumb and may b class-Z= mail. at that point the vote of finger is counted out of 100 arbitrary choices and similarly for the other two targets. In the event that finger is getting most noteworthy votes the last arbitrary woodland restores the finger as it anticipated objective. This idea is known as dominant part casting a ballot simply like races. Same applies to the remainder of the fingers of the hand. On the off chance that the calculation predicts the remainder of the fingers as fingers, at that point a significant level choice tree can cast a ballot that the picture is hand is a hand and this is the reason choice timberland is known as a gathering AI calculation.

Random Forest Pseudocode:

1. Accept the quantity of cases in the preparation set is N . At that point, the test of these N cases is taken aimlessly yet with substitution.
2. On the off chance that there are M information factors or highlights, a number $m < M$ is determined with the end goal that at every hub, m factors are chosen indiscriminately out of the M . The split on these m is utilized to part the hub. The estimation of m is held steady while we become the timberland.
3. Each tree is developed to the biggest degree conceivable and there is no pruning
4. Anticipate new information by amassing the forecast of the n tree trees (i.e., dominant part votes in favor of characterization, normal, for relapse).

Decision Tree

This is a sort of calculation which goes under the regulated learning procedure. It is a graphical portrayal of the multitude of potential answers for a choice. Choices depend on certain conditions. Choice made can be effortlessly clarified. Yet, first let me disclose to you why we pick the choice tree to assemble our task. Well, these choice trees are extremely simple to peruse and comprehend. It has a place with one of only a handful few models that are interpretable where you can see precisely why the classifier has settled on that specific choice. Little exertion needed for information arrangement. It can have both mathematical and all-out information. Non straight boundaries don't influence its exhibition. Choice trees start with the root and branches off to various arrangements simply like a tree. In the choice tree. Presently I will give a genuine situation

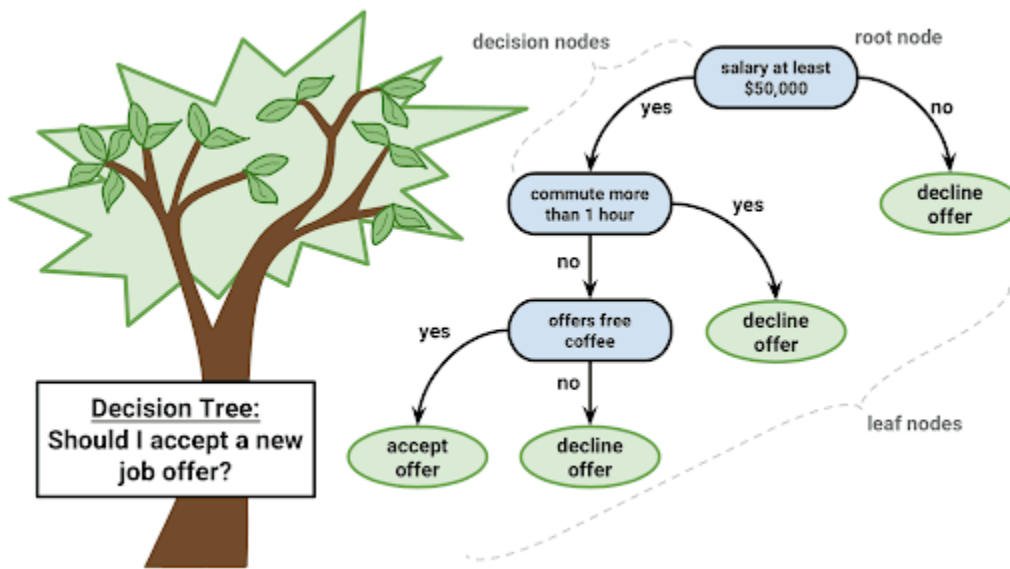


Fig 3.2.8: Real life scenario

On this specific picture we can see that the errand is should I acknowledge a new position offer or not. So you need to conclude that so u made a choice tree beginning with root hub was the essential compensation or least compensation is 50000 dollars in the event that it has not \$50000, at that point you can't acknowledge the offer okay at that point if your compensation is more prominent than \$50000 than you will additionally check if the drive is multiple in the event that it has offer than one hour you will decay the offer in the event that it is less, at that point one hour then you are drawing nearer to acknowledge the occupation then you will additionally check the organization is offering you a free espresso or not all that this is the case of choice tree.

Presently we will take care of the issue utilizing a choice tree classifier. Most importantly we need to make a dataset. We will include a few highlights information however with various level information. we need to separate it utilizing a Decision tree. The dataset is given beneath:

Table3.2.1: Dataset

COLOR	DIAMETER	LABEL
GREEN	3	Apple
YELLOW	3	Apple
RED	1	Grape
RED	1	Grape
YELLOW	3	Lemon

Here in the table we can see that the second and fifth model's width is the equivalent yet the name is unique so how about we perceive how the choice tree handles this issue.

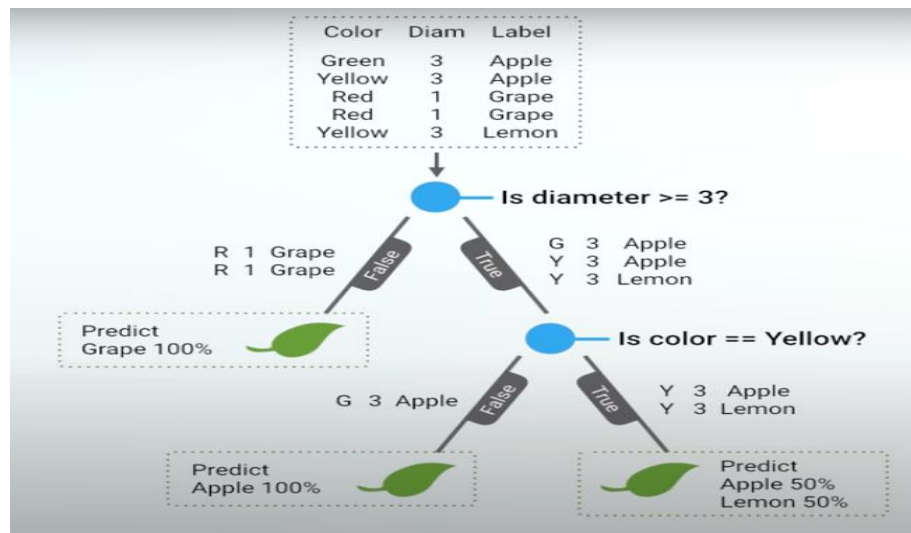


Fig3.2.9: How decision tree works

Over the figure we began with a root hub for the tree and all hubs get a rundown of lines as information. What's more, the root will get the whole preparing set now every hub will

pose a genuine bogus inquiry around one of the highlights and in light of this inquiry we split, or parcel, the information into two subsets. These subsets at that point become the contribution to two kid hubs we add to the tree. Also, the objective of the inquiry is to unmix the names as we continue down. Or then again at the end of the day, to deliver the most flawless conceivable conveyance of the marks at every hub. For instance, these hubs contain just a solitary sort of mark, so we would state it's completely unmixed. There's no vulnerability about this kind of name. Then again, the marks in this hub are as yet blended up, so marriage poses another inquiry to further restrict it down. What's more, the secret to building a viable tree is to comprehend which inquiry to pose and when. Furthermore, to do that we need to evaluate how much an inquiry unmixes the marks. At that point we can measure the measure of vulnerability at a solitary hub utilizing a measurement called Gini contamination. Furthermore, we cannot measure how much an inquiry decreases that vulnerability utilizing an idea called data pick up. We will utilize this to choose the best inquiry to pose at each point. Furthermore, given that question, we will recursively construct the tree on every one of the new hubs. We will keep partitioning the information and have no further inquiry to pose, at which pointer will add a leaf. To execute this, first we need to comprehend what sort of inquiry would we be able to pose about the information. Furthermore, second, we need to comprehend which inquiry to pose to when. Presently every hub accepts a rundown of columns as input. And to produce a rundown of inquiry we will emphasize over each incentive for each element that shows up in those rows. Each of these turns into a contender for an edge we can segment the information and there regularly be numerous potential outcomes. In light of an inquiry, we separation or parcel the information into Two subsets valid and bogus. The first contains all the lines for which the inquiry is valid. Also, second contains everything else. The best inquiry is the one that lessens our vulnerability the most. Gini pollution lets us measure how much vulnerability there is at a hub. Data gain will let us evaluate how much an inquiry diminishes that. Let's chip away at contamination first. Above all else take a lattice which ranges somewhere in the range of 0 and 1 where lower esteems demonstrate less vulnerability or blending at a hub. It measures our opportunity of being wrong on the off

chance that we haphazardly dole out a name from a set to a model in that set. Here is a model make that is understood.

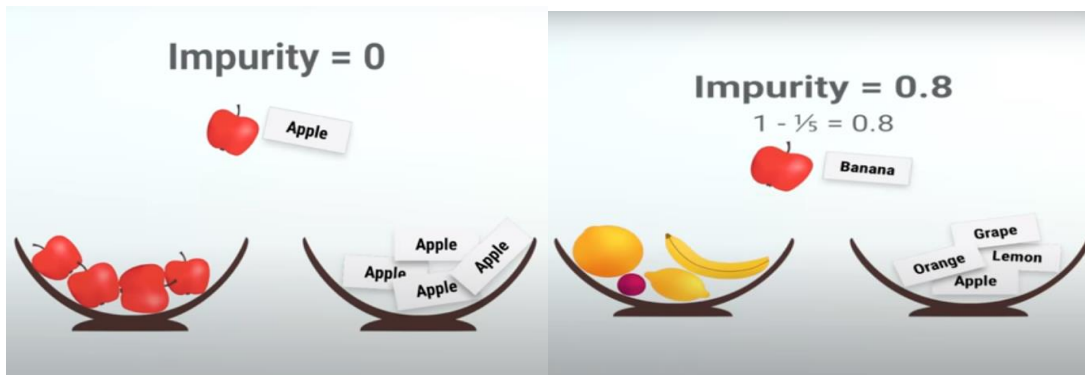


Fig3.2.10: Impurity

Envision we have two dishes one contains models and another contains marks. First, we will arbitrarily draw a model from the main bowl. At that point we arbitrarily draw a mark from the second. Presently we will order the model as having that name. Also, Gini contamination gives us our possibility of being off base. In this model we have just apples in each bowl there is no chance to commit an error so we state that the pollution is zero. On the other hand, given a bowled with five unique kinds of natural product I equivalent extent, we would state its debasement of 0.8. That is on the grounds that we have a one out of five possibility of being correct on the off chance that we arbitrarily name an example. Now data gain will let us discover the inquiry that diminishes our vulnerability the most. Furthermore, it's simply a number that depicts how much an inquiry serves to unmix the names at a hub.

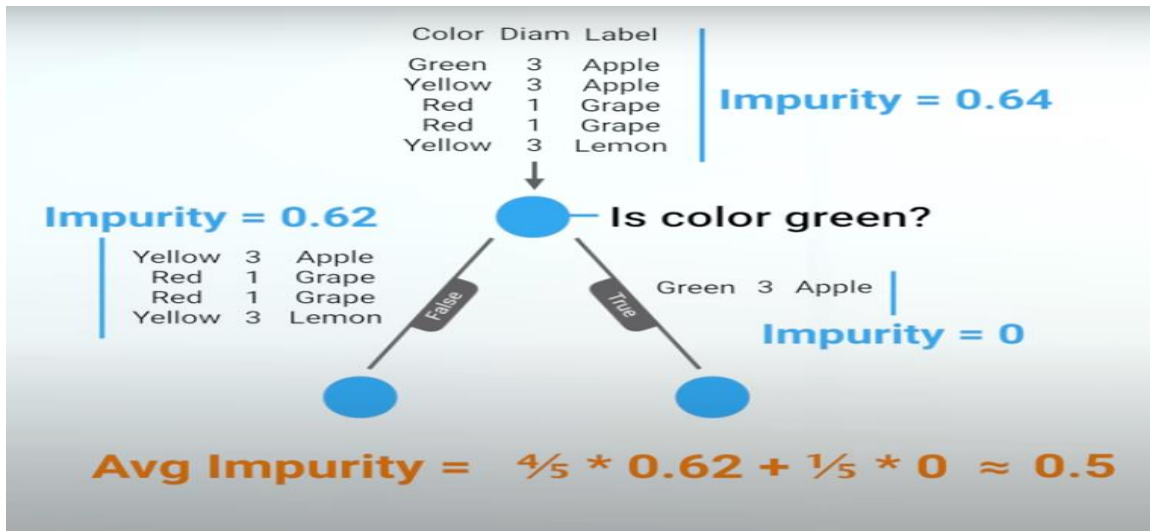


Fig 3.2.11: Avg Impurity Rule

Here is the thought we start by figuring the vulnerability of our beginning set in fig... .. At that point each inquiry we can pose to we will take a stab at parceling and computing the vulnerability of the youngster hubs that outcome. We will take a weighted normal of their vulnerability since we care more about an enormous set with low vulnerability then a little set with high. At that point we will take away this from our beginning uncertainty. And that is the data pick up and that is the way choice tree calculations work.

Support Vector Machine

It is a discriminative classifier that is officially planned by a separative hyperplane. It is a portrayal of models as focuses in space that are planned so the purposes of various classifications are isolated by holes as wide as could be expected under the circumstances. We use SVM on the grounds that it has some superb points of interest. It has high dimensional info space or some of the time alluded to as the scourge of dimensionality when we get to thousand measurements a ton of issues begin happening with most calculations that must be changed for the SVM consequently does it in high dimensional space one of the high dimensional space one high dimensional space that we take a shot at is inadequate report vectors this is the place where we tokenize the words in record so we can run our AI calculation over however I have seen ones get as high as 2.4 million unique

©Daffodil International University

tokens that is a ton of vectors to take a gander at lastly we have regularization boundary the acknowledgment boundary or lambda is a boundary that assists figure with trip whether we will have a predisposition or overfitting of the information whether it will be over fitted to a particular occasion or is going to be one-sided to a higher low an incentive with SVM it normally keeps away from the overfitting and inclination issues that we see in numerous different calculations these three favorable circumstances of SVM make it an integral asset to add your collection of AI apparatuses now . Why we use SVM lets give a genuine model: Lets we are searching for sweet strawberries and fresh apples in the store. We need to have the option to mark those two and choose what the organic product is and we do that by having information previously put in so

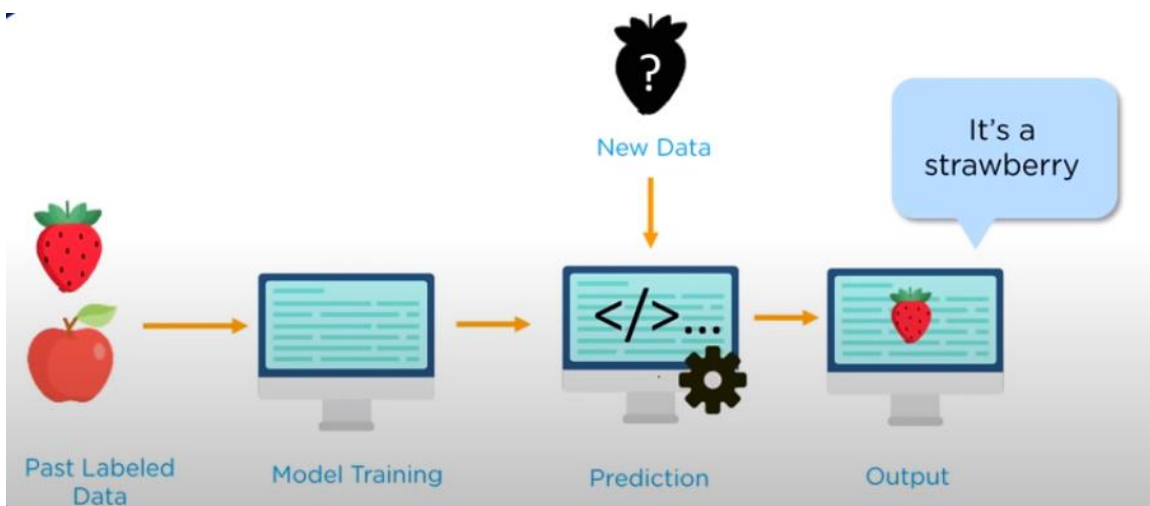


Fig 3.2.12: Example

we as of now have a lot of strawberries we know our strawberries and they are as of now named as such we as of now have a lot of apples we know our apples and are named as such then once we train our model at that point can be given the new information and the new information is this picture for this situation you see a question mark on it and it comes through and goes it's a strawberry for this situation, we are utilizing SVM model. SVM is a managed learning strategy that takes a gander at information and sorts it into one of two classes and for this situation we are arranging the strawberry into the strawberry site this

point you should be pose an inquiry how accomplishes the expectation work before we delve into an illustration of numbers how about we apply to our natural product situation.

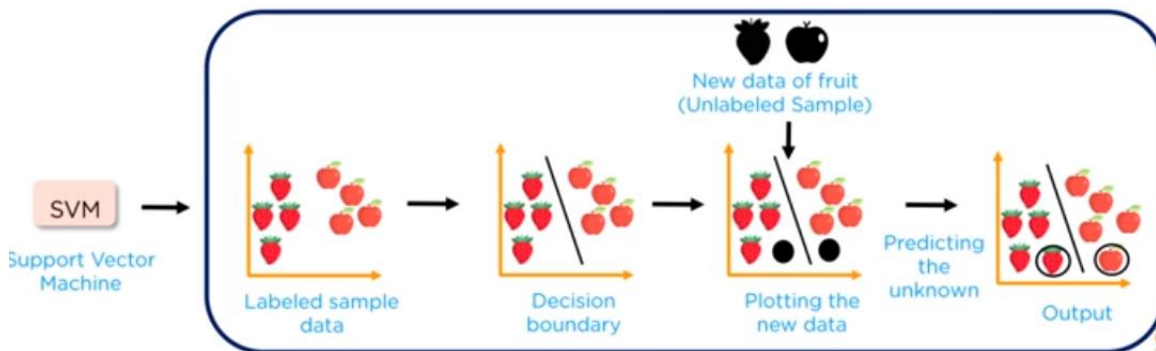


Fig 3.2.13: SVM entering

we have or SVM we have taken it and we take marked examples of information strawberries and apples and we draw a line down the center between the two gatherings. This split permits us new information for this situation: an apple and strawberry and spots them in the proper gathering dependent on which side of the line they fall in and that way we can foresee the accident as bright and scrumptious as a natural product model. Let's look at another model with certain numbers included and we would closer be able to take a gander at how the mathematical functions in this model.

Female		Male	
Height	Weight	Height	Weight
174	65	179	90
174	88	180	80
175	75	183	80
180	65	187	85
185	80	182	72

Fig 3.2.14: Sample Data Set

we are going to characterizing people and we are going to begin with a bunch of individuals with various stature and diverse weight and to make this work we should have an example informational index a female where you have their tallness weight 174 , 65, 174, 88and so on we will require an example informational collection of a mail they have stature 179,180 ... so on.Lets put this on a diagram so have a decent visual so you can see here we have two gatherings dependent on stature versus the weight on the left side we are going to have the ladies on the correct side we are going to have the men now in the event that we are going to make a classifier lets add another information point and sort it out if is male or female. Thus, before we can do that we need to part to our information first we can part our information by picking any of these lines for this situation we attracted two lines through the information the center that isolates the men from the ladies yet to anticipate the sex of another information point we should part the information in the most ideal manner and we

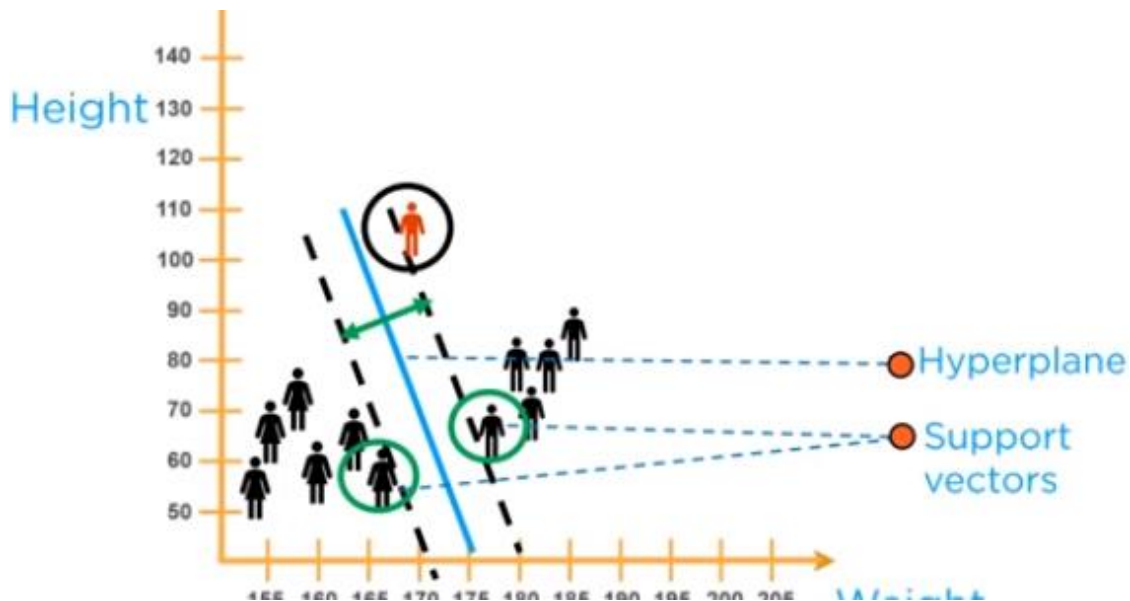


Fig3.2.15: SVM graph

state the most ideal way this line has a greatest space that isolates the two classes here you can see there's a reasonable split between the two distinct classes and in this one there's less an unmistakable split this doesn't have the greatest space that isolates the two that is the reason this line best parts the information. We would prefer not to simply do this by eyeballing it and before we go further, we need to add some specialized terms to this. We

©Daffodil International University 27

can likewise say that the distance between the focuses and lines should be quite far. In specialized terms we can say the distance the help vector and hyperplane should be quite far , and this is the place where the help vectors are the extraordinary point in the informational collection and on the off chance that you take a gander at this informational collection they have surrounded two focuses which is by all accounts directly on the edges of the ladies and one of the edges of the men and hyperplane has a greatest distance to the help vector of any class now you will see the line down the center . We consider this the hyperplane on the grounds that when you are managing numerous measurements it truly isn't only a line yet a plane of crossing points. you can see here where uphold vectors have attracted run lines. The math behind this is exceptionally basic: we take D_+ the briefest distance to the nearest sure point which would be men's side and D_- is the most limited distance to the nearest negative point which is on the ladies' side. The amount of D_+ and D_- is known as the distance edge or the distance between the two help vectors that appeared in run lines and afterward by finding the distance edge, we can get the ideal hyperplane. When we make an ideal hyperplane, we can without much of a stretch see which new information fits in and dependent on the hyperplane we can say that the new information point has a place with the male sex ideally how that chips away at a visual level as an information. There is another inquiry that can be posed. What occurs if the hyperplane isn't ideal? In the event that we hyperplane having a low edge, at that point there is high possibility of miss characterization. This specific SVM model the one we talked about so far is likewise called or alluded to as the SVM

3.3 Proposed Model

A model is proposed when offered and a contention the component of the informational index model and likewise talked about. We have proposed a model which gets the contribution to the message set. This content set has some assessment as audit or remark which is utilized in the examination for taking assumptions. Two significant layers in our proposed model one is information handling and another is dissecting slant. Information handling layers talk about the assortment of information, pre-preparing information and distraught favor for conclusion examination. Another layer measures the information for the train and testing in a strategy and gets the normal yield. We will be talked about quickly in the accompanying segment.

A flowchart is a graph that portrays a cycle, framework or PC calculation. They are broadly utilized in numerous fields to archive, study, plan, improve and impart frequently complex cycles in clear, straightforward charts.

Let's have a quick look of my working process flowchart I given below:

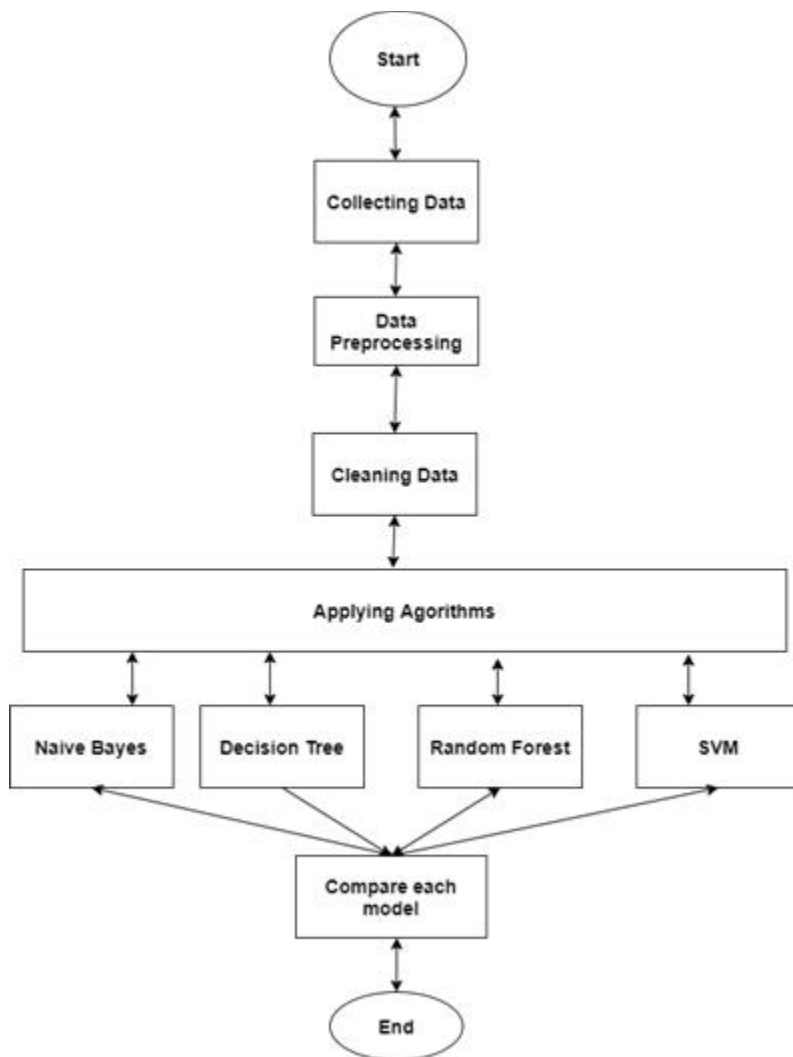


Fig 3.3.1: Working Process flowchart

3.4 Data Collection Procedure

To gather information, we created a google docx where our friends and families help us to get some great data. We search every social media post which is related to religion. It takes so much time to collect data from social media. We collect data from those persons who are very faithful to their religion and we also collect data from atheists. We collect data from various websites and from quotes about religion.

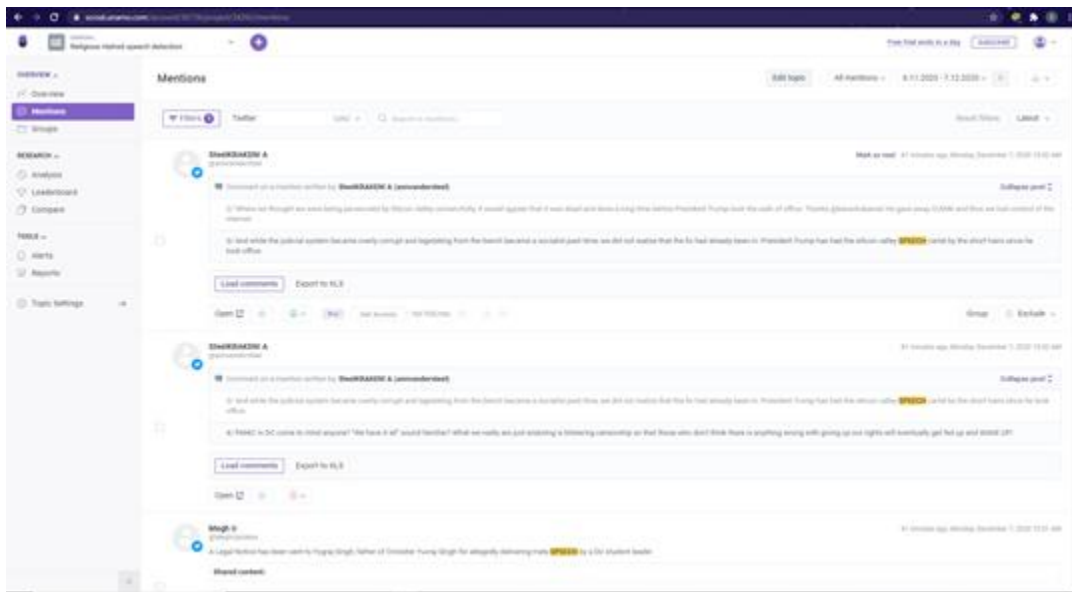


Fig 3.4.1: Collecting data from social media

content	sentiment
1 content	
2 Idolatry is Forbidden.	Negative
3 Islam is a peaceful Religion.	Positive
4 Fast is important for body.	Positive
5 Charity comes with poverty.	Negative
6 Salah is good for health.	Positive
7 Every religion believes in peace.	Positive
8 Zakat can remove poverty.	Positive
9 God lives within every creation.	Positive
10 Jewish makes quarrel.	Negative
11 Every muslim is not terrorist.	Positive
12 God lives in top floor.	Negative
13 Muslims are terrorist.	Negative
14 Hindus worship cow.	Negative
15 Hazrat Muhammad (S) is the best creation of Allah.	Positive
16 Women deserve to be abused.	Negative
17 Muslims want war.	Negative
18 Islam is the true religion. believe it brothers.	Positive
19 Kill those pig eaters fuck those americans send them to hell.	Negative
20 Christians are all demonic.	Negative
21 Buddhism is a religion for cows. People who believe this shit they get chopped sooner or later.	Negative
22 Jews are lower class pigs.	Positive
23 Macca is the best place for muslims.	Positive
24 Jins exist.	Positive

Fig 3.4.2: Data collection and store

The ratio of Positive and Negative Data in our Dataset is given below:

Positive Negative

48.96 51.04

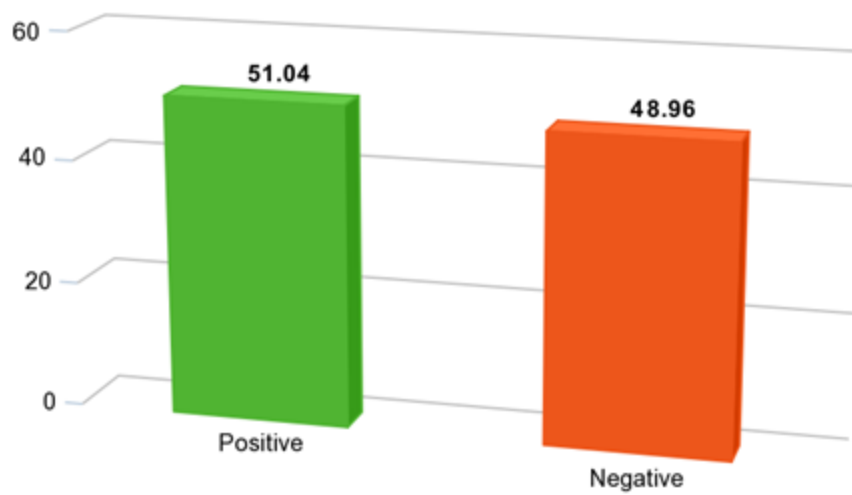


Fig 3.4.3: Percentage of positive and negative data

We have to work on some preprocessing of our dataset. We need to convert all our data into lowercase, then we remove square brackets, numbers and punctuation.

Table 3.4.1: Data cleaning process

	Content	Sentiment	Cleaned Content
1	Idolatry is Forbidden.	Negative	idolatry is forbidden
2	Islam is a peaceful Religion.	Positive	islam is a peaceful religion
3	Fasting is important for body.	Positive	fasting is important for body
4	Charity comes with poverty.	Negative	charity comes with poverty
5	Salah is good for health.	Positive	salah is good for health

3.5 Implementation and Requirements

In this paper we are using 2 types of string data. Positive and Negative speech about Religion. We collect two types of dataset from different sources. We went to those people who obey their religion very much. We hear from them and collect data from their speech. Even we went to atheists and also listen to them. We also analyze their speech and try to collect data from them too. We try to collect data from websites, social media like Facebook, Twitter, YouTube etc. All of the datasets have been separated into training and testing sets with individual subsets. Here we use four different types of algorithms to get our maximum accuracy.

CHAPTER 4

Experimental Results and Discussions

4.1 Introduction

The extended model has taken an exceptional kinds of AI algorithmic guideline procedures to improve the exactness and examination that one is gives us higher precision. This method had a go at depleting it for higher precision. The exactness of the model was determined that depict in given beneath.

4.2 Experiment Process

We used four NLP techniques they are Naive Bayes, Random Forest, Decision Tree and SVM.

We do the following things here,

- Collect Data
- Data Pre-Processing
- Data Cleaning
- Apply Classifier
- Comparison between Classifier
- Find the best accuracy

The confusion matrix we get from **Naïve Bayes** is given below:

```
Array ([[23, 14],  
       [ 9, 31]])
```

TP_NB = 23 #True Positives

TN_NB = 31 #True Negatives

FP_NB = 14 #False Positives

FN_NB = 9 #False Negatives

The result we get from the matrix,

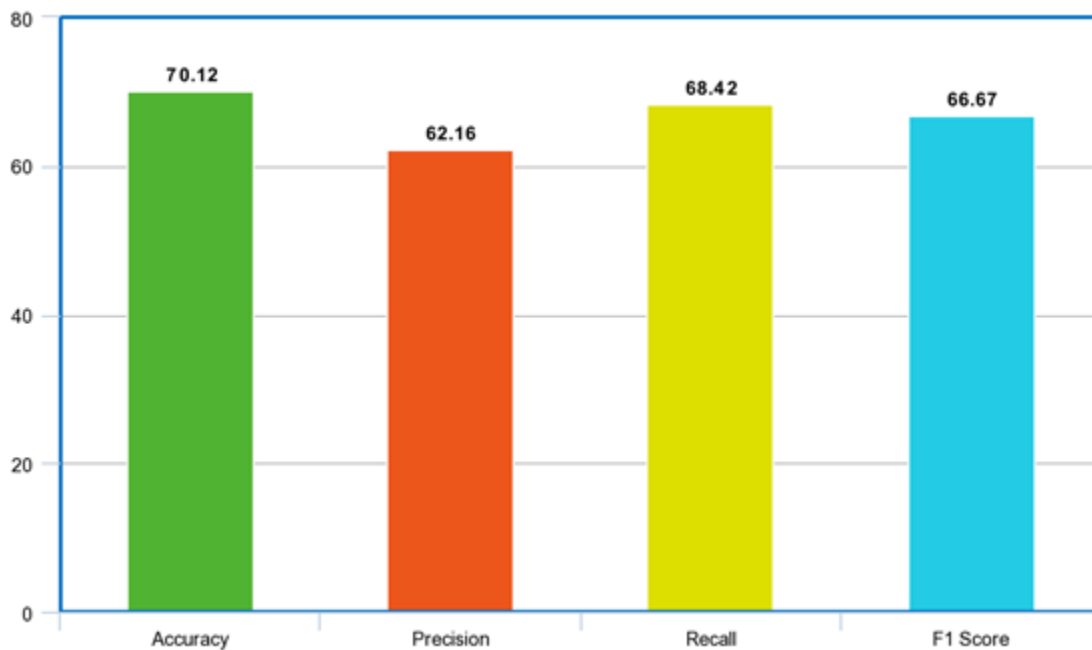


Fig 4.2.1: Value of Naïve Bayes Classifier

The Confusion Matrix we get from **Random Forest** is given below:

```
Array ([[32, 5],  
       [16, 24]])
```

TP_RF = 32 #True Positives

TN_RF = 24 #True Negatives

FP_RF = 5 #False Positives

FN_RF = 16 #False Negatives

The result we get from the matrix,

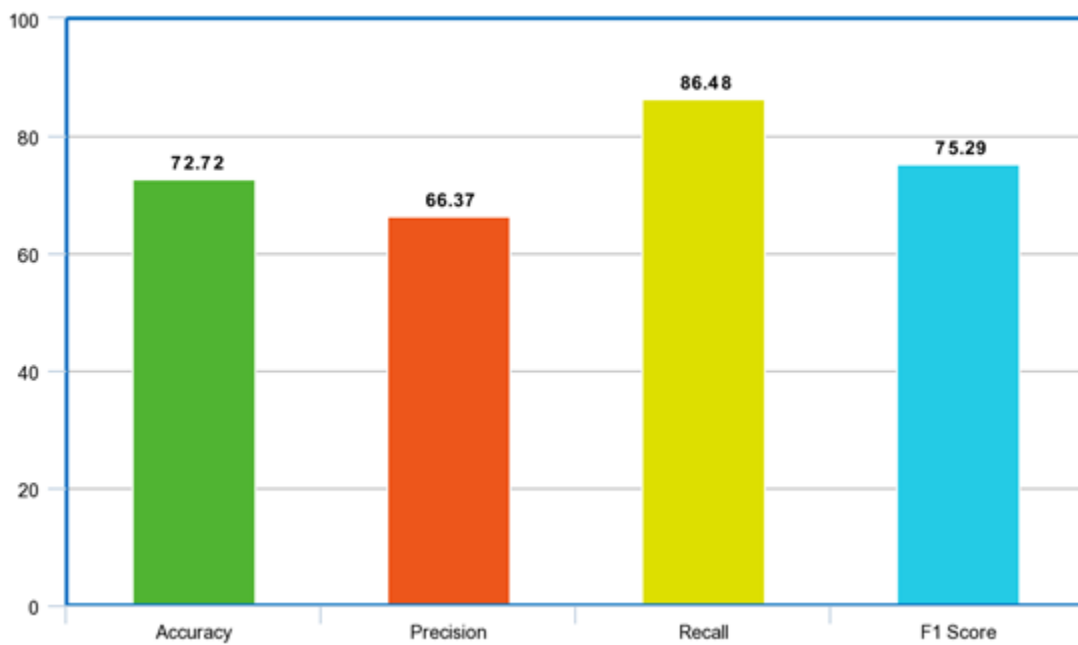


Fig 4.2.2: Value of Random Forest Classifier

We also get another Confusion Matrix from **Decision Tree**.

```
Array ([[24, 13],  
       [14, 26]])
```

TP_DT = 23 #True Positives

TN_DT = 26 #True Negatives

FP_DT = 13 #False Positives

FN_DT = 14 #False Negatives

The result we get from the Confusion Matrix,

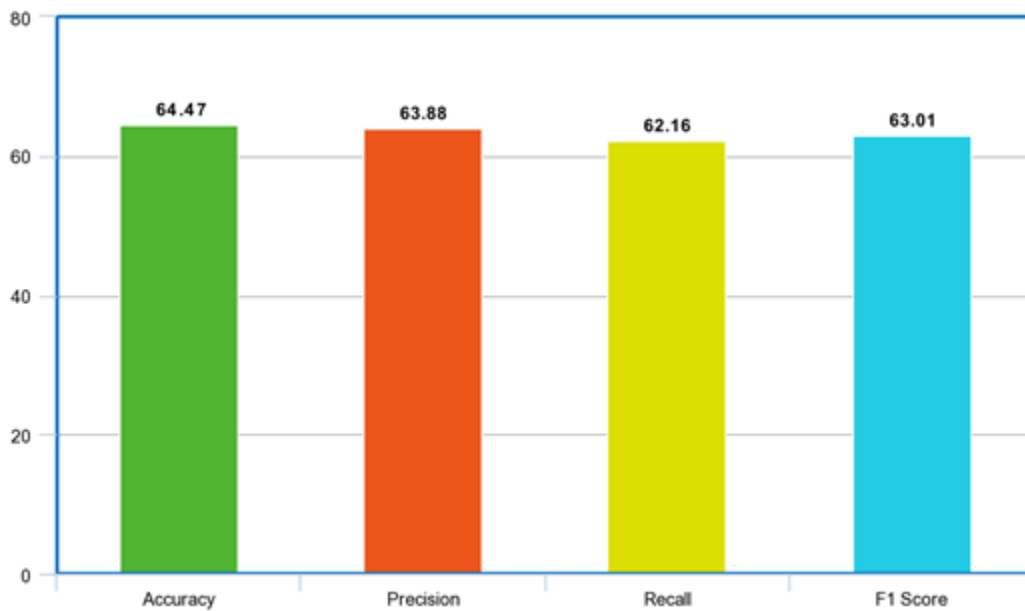


Fig 4.2.3: Value of Decision tree Classifier

The Confusion Matrix we get from **SVM** is given below:

```
Array ([[32, 5],  
       [16,24]])
```

TP_SVM = 32 #True Positives

TN_SVM = 24 #True Negatives

FP_SVM = 5 #False Positives

FN_SVM = 16 #False Negatives

The result we get from the confusion matrix,

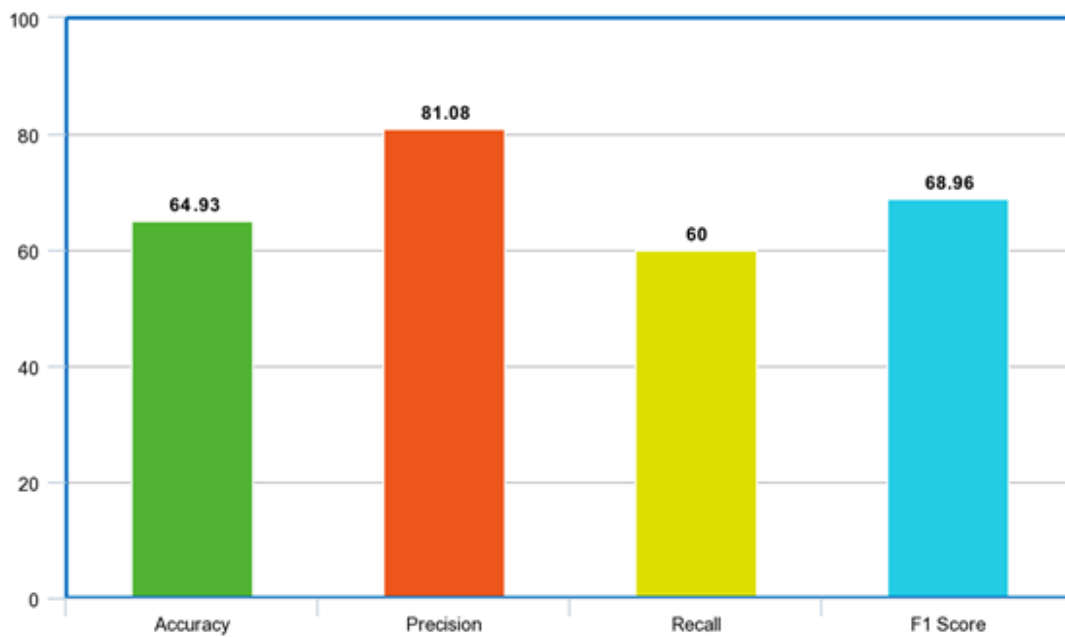


Fig 4.2.4: Value of SVM classifier

4.3 Experimental Result

Table 4.3.1: The results in different classifier

Classifier	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	70.13	62.16	68.42	66.67
Decision Tree	64.47	63.88	62.16	63.01
Random Forest	72.72	66.67	86.48	75.29
SVM	64.93	81.08	60.00	68.96

We can see the table where we get different results for different kinds of classifiers. We get the best accuracy in Random forest classifier which is 72.72%, The best precision we get from is SVM which is 81.08%, Random forest gives us best recall rate which is 86.48% and F1 Score is also highest in Random forest which is 75.29%.

Now look at the comparison diagram,

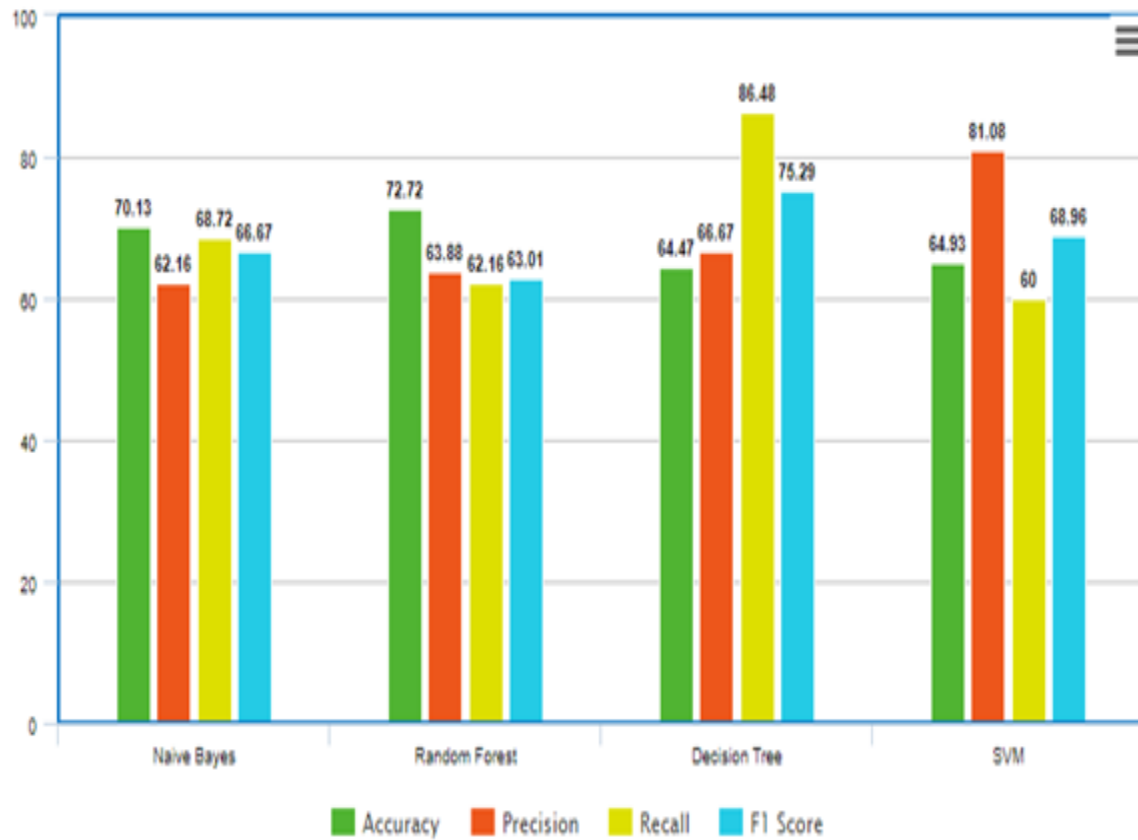


Fig 4.3.1: Comparison of different classifier

4.4 Summary

In this chapter we find out our best accuracy, precision, recall and F1 score by using the table and the comparison diagram. And we briefly discussed our result and gave a summarized explanation.

CHAPTER 5

Conclusion and Future Scope

5.1 Conclusions

Sentiment Analysis is a well-known theme for information examination and announcing and progressed handling. In this examination primary objective is to speak to the diverse kinds of classifiers and think about their outcome when arranging kinds of surveys, all things considered. The proposed approach should be more exact in the event that we change the boundary in the classifier. From all of our four classifiers Random forest gives the best accuracy. So, we can say that we should use the Random forest classifier to implement our project.

5.2 Future Scope

In our project we have worked with Natural Language Processing (NLP) techniques to find out extreme religious discourses from religious comment and post on social media. NLP techniques have different classifiers but for now we have used Naïve Bayes, Random Forest, Decision Tree and SVM classifiers. In future we will work with more data and use other algorithms of Natural Language Processing techniques and try to increase the accuracy of classification. It will help the cybercrime security authorities to find the religious violence related content. We hope it can measure the effectiveness of religious violence. Our work will help to concern the people about the abuse of online sites.

References

- [1] Abdelkader Rhouati, Jamal Berrich, Mohammed G. Belkasmi, & Toumi Bouchentouf. (19-6-2018). Sentiment Analysis of French Tweets based on Subjective Lexicon Approach: Evaluation of the use of OpenNLP and CoreNLP Tools. *Team SIQL, Laboratory LSEII, ENSAO Mohammed First University, Oujda, Morocco*, 830-836.
- [2] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, & Dr.M.Venkatesan. (Dec 2015-Jan 2016). Twitter Sentiment Analysis of Movie Reviews using Machine Learning. *Akshay Amolik et al. / International Journal of Engineering and Technology (IJET)*, 7, 2038-2044.
- [3] Akshma Chadha, & Baijnath Kaushik. (3 September 2019). A Survey on Prediction of Suicidal Ideation Using Machine and Ensemble Learning. *Section C: Computational Intelligence, Machine Learning and Data Analytics*, 1-16.
- [4] Anna Baj-Rogowska. (2017). Sentiment Analysis of Facebook Posts: the Uber case. *The 8th IEEE International Conference on Intelligent Computing and Information Systems (ICICIS 2017)*, 391-395.
- [5] Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, & Jaime Caro. (July 2013). Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning.
- [6] Desmond U. Patton, William R. Frey, Kyle A. McGregor, Fei-Tzin Lee, Kathleen McKeown, & Emanuel Moss. (February 2020). Contextual Analysis of Social

Media: The Promise and Challenge Of Eliciting Context in Social Media Posts with NaturalLanguage Processing.

[7] Efthymios Kouloumpis, Theresa Wilson, & Johanna Moore. (n.d.). Twitter Sentiment Analysis:The Good the Bad and the OMG! 538-541.

[8] Ilham Safeek, & Muhammad Rifthy Kalideen. (December 2017).
PREPROCESSING ON FACEBOOK DATA FOR SENTIMENT ANALYSIS. *Proceedings of 7th International Symposium, SEUSL, 7th & 8th December 2017*, 69-78.

[9] Jibon Naher, & Matiur Rahman Minar. (2016). Impact of Social Media Posts in Real life Violence: A Case Study.

[10] Khaled Ahmed, Neamat El Tazi, & Ahmad Hany Hossny. (September 2015).
Sentiment Analysis Over Social Networks: AnOverview.

[11] R. A. S. C Jayasanka, M. D. T. Madushani, E. R. Marcus, I. A. A. U. Aberathne, & s. c. premaratne. (n.d.).

[12] Taimur Islam, Ataur Rahman Bappy, Tanzila Rahman, & Mohammad Shorif Uddin. (2019). Filtering Political Sentiment in Social Media from. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 663-666.

[13] Berger, A. L., Vincent, J. D., and Stephen, A. D. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1): 39-71.

[14] Das, S., and Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. 8th Asia Pacific Finance Association Annual Conference (APFA).

[15] Hemnath, R., and Low, B. W. 2010. Sentiment Analysis Using Maximum Entropy and Support Vector Machine. *Semantic Technology and Knowledge Engineering in 2010*. Kuching, Sarawak.

[16] Ibrahim, A. E.-K. 2006. Effect of Stop Words Elimination For Arabic Information Retrieval: A Comparative Study. International Journal of Computing & Information Sciences, 119-133