Shahjalal University of Science and Technology
Department of Computing Science and Engineering

CSE 408

A Naive Bayes Approach to Identify Users Gender, Age and Religion Based on their Facebook Snippets

Md. Abdullah Al Awal

Reg. No.: 2012331055

4^{th} year, 2^{nd} Semester

Farzad Bin Fazle

Reg. No.: 2012331005

4^{th} year, 2^{nd} Semester

Department of Computer Science and Engineering

Supervisor

Md. Saiful Islam

Assistant Professor

Department of Computer Science and Engineering

17th September, 2017

A Naive Bayes Approach to Identify Users Gender, Age and Religion Based on their Facebook Snippets

A Thesis submitted to the

Department of Computing Science and Engineering

Shahjalal University of Science and Technology

Sylhet - 3114, Bangladesh

in partial fulfillment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering

By

Md. Abdullah Al Awal

Reg. No.: 2012331055

4^{th} year, 2^{nd} Semester

Farzad Bin Fazle

Reg. No.: 2012331005

4^{th} year, 2^{nd} Semester

Department of Computer Science and Engineering

Supervisor

Md. Saiful Islam

Assistant Professor

Department of Computer Science and Engineering

17th September, 2017

Recommendation Letter from Thesis Supervisor

The thesis

entitled " A Naive Bayes Approach to Identify Users Gender, Age and Religion Based on their Facebook Snippets"

submitted by the students

1. Md. Abdullah Al Awal
2. Farzad Bin Fazle

is a record of research work carried out under my supervision and I, hereby, approve that the report be submitted in partial fulfillment of the requirements for the award of their Bachelor Degrees.

Signature of the Supervisor:

Name of the Supervisor: Md. Saiful Islam

Date: 17th September, 2017

Certificate of Acceptance of the Thesis

The thesis

entitled "A Naive Bayes Approach to Identify Users Gender, Age and Religion Based on their Facebook Snippets"

submitted by the students

1. Md. Abdullah Al Awal
2. Farzad Bin Fazle

on 17th September, 2017

is, hereby, accepted as the partial fulfillment of the requirements for the award of their Bachelor Degrees.

Acting Head of the Dept.	Chairman, Exam. Committee	Supervisor
M. Jahirul Islam, PhD, PEng	Dr Mohammad Reza Selim	Md. Saiful Islam
Professor	Professor	Assistant Professor
Department of Computer Science and Engineering	Department of Computer Science and Engineering	Department of Computer Science and Engineering

Abstract

Is there lying any actual difference in their posts between men and women or people of different religion or age? Today it is very important to know that whether a text is written by male or female or is it possible to identify a persons gender by using their writing? Several work described a lot differences between men and women and people of different religion. Besides it is proved that there is actually linguistics, psychological and sociological difference between men and women or difference between people of different religion. In this paper, we tried to find out the dissimilarities between them using several text classification algorithms, analyze performance of those algorithms and reports their results on Facebook posts and comments.

Keywords: Classification, Facebook, Gaussian Naive Bayes, SVM, Neural Network, Linear Regression, Logistic Regression.

Acknowledgements

We would like to thank the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh, for supporting this research. We are very thankful to our honorable teacher Md. Saiful Islam for his outstanding directions.

We are also grateful to anonymous authors of previous works for their co-operation and support. We want to give an special thanks to our friend Quazi Ishtiaque Mahmud Rafi and Bishwajit Purkaystha for their helping hand.

Dedication

We would like to dedicate our research to our parents. We are also grateful to anonymous authors of previous works for their co-operation and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today social media is a great source of data. Everyday people are using these sites and leaving out a huge amount of data about their current situation, what they likes, their photos, opinions, feelings, places they visited, among their information. This information much more relevant because there is no difference between real world and social world. People express themselves in real world as they express themselves in SNS. So predicting users gender using Facebook status is a best idea. Main reason for that is the huge amount of relevant data. SNS like Facebook, Twitter, Weibo etc. are popular right now. LinkedIn is already transformed themselves into some kind of SNS. Facebook is one of the most popular types of SNS. It has 1.86 billion monthly active users, as of the fourth quarter of 2016. There are about 3.7 billion internet users (about 47% of world population) and about 2.3 billion use social media.

People are providing huge amount of data in their day to day life. For example, number of likes in a status, number of tag which defines how social that specific user is, number of friends, birthday, number of group managed, number of cover photo, profile pictures and albums, check-in, number of status, city name etc. These features contains a lot of information about that specific user.

That's why in our approach we considered users Facebook status. Status can vary from gender to gender. There are a lot of similarity in status between male and female. But less difference between them. We tried to find dissimilarity between them with our model, test our model and showed that it is possible to classify gender according to their status.

Similar task was done on age features. Status can also vary from age to age. Person with different

age upload different types of status. So it is also possible to classify the age level depending on their age.

Religion can also be classified by status. Person with different religion upload different religious status. Not every people upload religious views in social media. So sometimes it is difficult to identify their religious views. So we tried to keep balance in religious ratio. If the ratio will maintained well our model will also be performed well.

1.1 Motivation

In our first research work, we first studied to find the user personality through their social network site. Several works are done on personality measurement. Then we turn back and started to work on user classification like Gender (male/female) identification, Age and Religion classification. We have been motivated that there are more similarity on Facebook post snippets between Male and Female. And less difference between them. That's why we move forward and started working on users age, gender and religion classification through their Facebook post snippets and got better performance on some features.

Chapter 2

Background Study

2.1 Literature Review

In our study we found a few amount of academic approved papers related to our task personality prediction. Since the increasing information available on social networks, many authors interested in trying to predict userâs personality with the help of information people shared in Facebook.

Authors of [1] accurately measured the user personality through the public information available on their Facebook profiles along with 44 items and Big Five dimensions. Some of them are number of photos, features like relationship, date of birth, status, religion, education history, gender, hometown, personal activities etc. Authors tried to find the correlation between the data and Big Five Inventory and used profile data as feature and trained those data with two machine learning algorithm called m5sup/Rules and Gaussian Processes. Authors found a good result with RMSE error 0.73 for Openness, 0.73 for Conscientiousness, 0.99 for Extraversion, 0.73 for Agreeableness and 0.83 for Neuroticism.

Another work done on user personality prediction was described in [2]. Authors used both the Facebook userâs profile extracted data and also collected each data by Curiosity Exploration Inventory (CEI-II) form. For each user they collected total number of group they involved, total number of photos, total number of friends, total number of likes and some basic profile features like primary, high school and university (CEI-II) degree etc. Then they try to find correlation between the features and degree of curiosity. Finally they proposed a model with a decision tree with 3 searches and 8 evaluator methods. It is very important to notify that they used Weka3 an

open source software to check performance their model. And showed that it is possible to predict a personâs curiosity with some specific features using J48 algorithm with a 10-cross-validation.

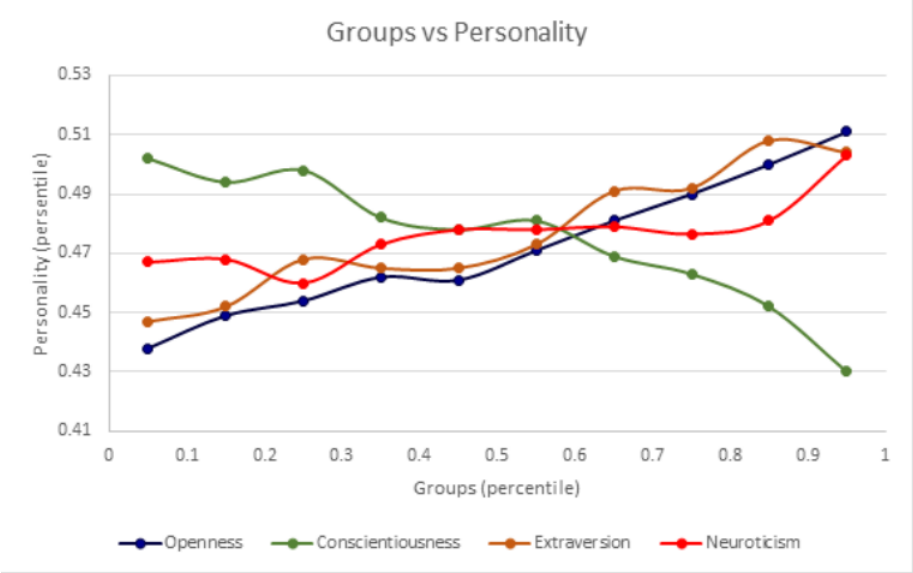

Figure 2.1: Groups vs Personality (percentile)

In another work author extracted 1,80,000 userâs Facebook profile containing number of photos, number of groups, number of likes, size and density of friendships etc. Then tried to find correlation between their personality and profiles. It was showed that the best accuracy achieved for Extraversion and Neuroticism, the lowest accuracy achieved for Agreeableness and middle accuracy for Openness and Conscientiousness [Table: 2.1]. In their methodology they collected all Facebook features and they sorted the n users according to their number of friends feature for example from the user with the smallest number of Facebook friends to the user with the greatest number of Facebook friends, to obtain the sorted list u1, u2,....., un. They denote each userâs number of friends as ci. Then they partitioned those ordered users into k equal and disjoint sets. Where the set S1 of q = n/k users with the smallest number of friends), the following set S2 of q users with slightly higher feature values and so on until the set Sk contains q users of the highest feature values(users with the most friends). Authors partitioned users into k=10 large groups. Then they plotted Clustered Scatter Plots to show the relationship between Big Five Inventory Model and Facebook features where horizontal axis represented the average Facebook feature value and vertical axis represented the average personality trait score [Fig-2.1,2.2].

Trait	R^2 (%)	RMSE (%)
Openness	0.11	0.29
Conscientiousness	0.17	0.28
Extraversion	0.33	0.27
Agreeableness	0.01	0.29
Neuroticism	0.26	0.28

Table 2.1: Predicting personality traits using Facebook features through multivariate linear regression

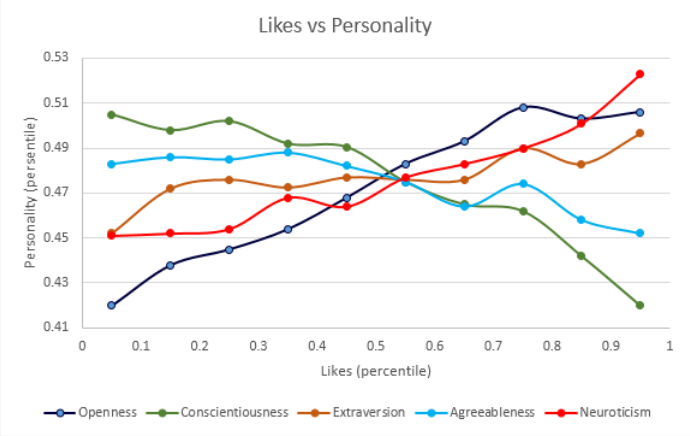

Figure 2.2: Likes vs Personality (percentile)

Another work described in [4]. Authors crawled 479k users profile data and analyzed based on three categories like gender, study that depicts the age distribution of users and inferred data from Facebook userâs profile. Mainly authors focused on the closed and disclosed section and measure degree of each Facebook userâs profile. Finally they showed that closed and disclosed section of users profile is depending on their age distribution [Fig-2.3] and gender distribution [Table-2.2].

Another work described in [5] authors collected users profile attributes, interests, advertising content users are exposed to is either relevant or irrelevant or not possible to explain based on their online activities [Table 2.3]. Then authors first applied Singular Value Decomposition (SVD) with k (user)-dimensional vector to build the model and used that model to for some unknown users to train.

After analyzing authors focused on four target label pairs of sensitive nature a) gay vs straight, b) single vs married, c) liberal vs conservative and d) Christian vs Muslim. Then authors run feature selection process and removed those users that did not contain any of above label. Then

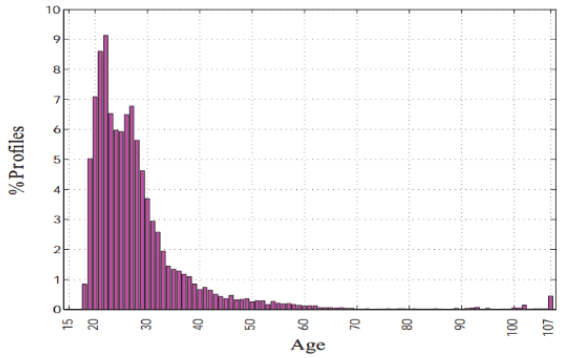

Figure 2.3: % Profiles in different age range

Attributes categories	%Male (%)	%Female (%)
All	51.33	48.67
Friend-list	53.99	46.01
CurrentCity	52.81	47.19
HomeTown	54.05	45.95
Gender	51.33	48.67
Birthday	49.23	50.77
Employers	55.23	44.77
College	53.30	46.70
HighSchool	55.89	44.11

Table 2.2: Gender analysis per categories of attributes

authors showed that provide a clue that gay and straight can be performed more accurately than single and married.

In [6], authors tried to differentiate the relative influences of different feature extraction approaches like information retrieval feature (TFIDF), document frequency (df), linguistic feature(lin), length of social snippet(lenText), relative position(pos) etc. on social snippets and showed that TFIDF has most. Figure 2.4 shows their analysis.

L	Labels	Users(n)	Balance	m	m′
L1	Gay/Straight	2412	50/50	218490	15609
L2	Single/Married	7732	50/50	511775	29389
L3	Liberal/Conserv.	4106	55/45	296298	18658
L4	Muslim/Christian	1196	74/26	134120	10333

Table 2.3: Dataset overview

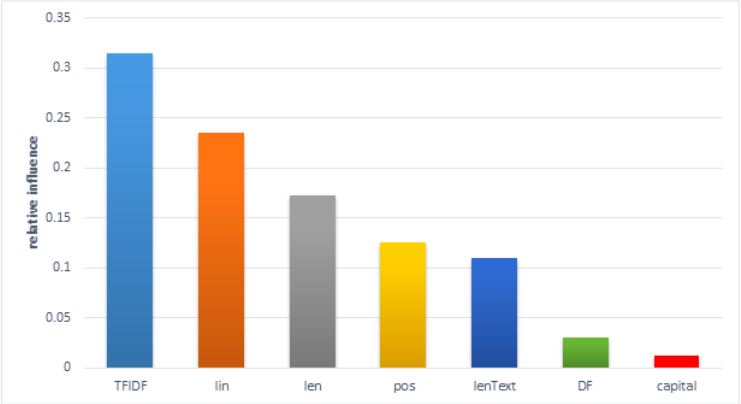

Figure 2.4: The relative influence of each feature in Gradient Boosting Machine (GBM)

Chapter 3

Methodology

In this chapter we will discuss different classification techniques. Several approaches were applied in previous work. Some were perform with best accuracy, some were bad accuracy and take several times. We will discuss various classification techniques and feature extraction techniques.

3.1 Term Frequency

Tf means term-frequency. Term frequency defines by tf(t,d). The easiest way is to use the actual count of a term t in a document for example, if the number of times that a term t occurs in document d and if we denote the actual count by (ft,d), then the simplest term frequency scheme is:

$$tf(t,d) = (f_{t,d})$$

3.2 Inverse Document Frequency

Idf means inverse document-frequency. It denotes number of times of a term t that contains in a given document is multiplied with idf. There are several formula to calculate idf. They are slightly different from each other. Below is one of them:

$$\text{Idf(t)=log } \frac{n_d}{1+df(d,t)}$$

Here total number of documents is denoted by n_d and df(d,t) is denoting the number of documents that contains the term t. The tf-idf vectors are then normalized by the below Euclidean norm:

$$v_{norm} = \frac{v}{||v||^2} = \frac{v}{\sqrt{{v_1}^2+{v_2}^2+\cdots+{v_n}^2}}$$

3.3 Term Frequency Inverse Document Frequency

As we defined before the definition of tf-idf. To calculate tf-idf of a given corpus we need to calculate tf and idf individually and multiplied both. Simple formula of tf-idf calculation is:

$$\text{tf-idf(t,d)=tf(t,d)} \times \text{idf(t)}$$

We can calculate idf in several way though they are slightly different from each other as stated before. TfidfTransformer and TfidfVectorizer. We used TfidfVectorizer in our experiment.

3.4 Gaussian Naive Bias

Gaussian Naive Bayes is a simple probabilistic classifier that uses Bayes theorem with independent assumptions between the features to classify data. Let, the classifier has m elements which is denoted with $X = X_1, X_2, \ldots, X_m$ and n classes which is denoted with $C = c_1, c_2, \ldots, c_n$. The Bayes theorem is stated in following equation,

$$P(C_i \mid X_j) = \frac{P(X_j|C_i)*P(X_i)}{P(X_j)}$$

Here,

C_i = Denotes the class

X_j = Denotes a single featured element

P(A | B) = Denotes the probability of observing A after B is observed.

P(A) = Denotes the probability of observing A

For multiple feature the equation is changed to,

$$P(C_i \mid x_1,x_2,x_3, \ldots , x_n) = \frac{P(x_1|C_i)*P(x_2|C_i)*\ldots*P(x_n|C_i)}{P(X_j)}$$

Here,

$x_1,x_2,x_3, \ldots , x_n$ are features of X_j.

Gaussian Naive Bayes is really fragile to over fitting without any regularization assumption. Also, it is based on naive assumptions that re not generally concordant with the data.

3.5 Support Vector Machine

SVM is a supervised learning algorithm which is used analyze classification and regression analysis data. SVM is normally used to classify data into 2 categories. It performs better than Gaussian Naive Bayes. In SVM, the training data is plotted in the hyperspace and a hyper plane is drawn which separates the data into 2 categories as widely as possible. This hyper plane is called SVM.

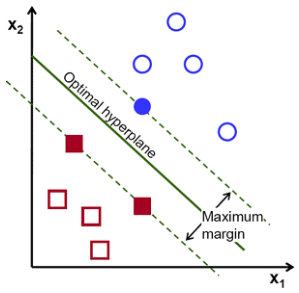

Figure 3.1: Basic Support Vector Machine

If there exists multiple class in the training data set, then multiple SVM is drawn to classify them. If the number of feature is too large or number of data set is too large then SVM does not perform well.

3.6 Linear Regression

Linear regression is a linear approach to compute the relationship between the dependent variable and the independent variables. Linear regression is often used in predictive analysis. In linear regression, a straight line is drawn which follows the below equation and known as hypothesis function.

$$h_\theta(x) = \theta^T x$$

Here,

θ = Denotes co-efficient matrix

X = Denotes the matrix of independent variables

$h_\theta(x)$ = Denotes a hypothesis function.

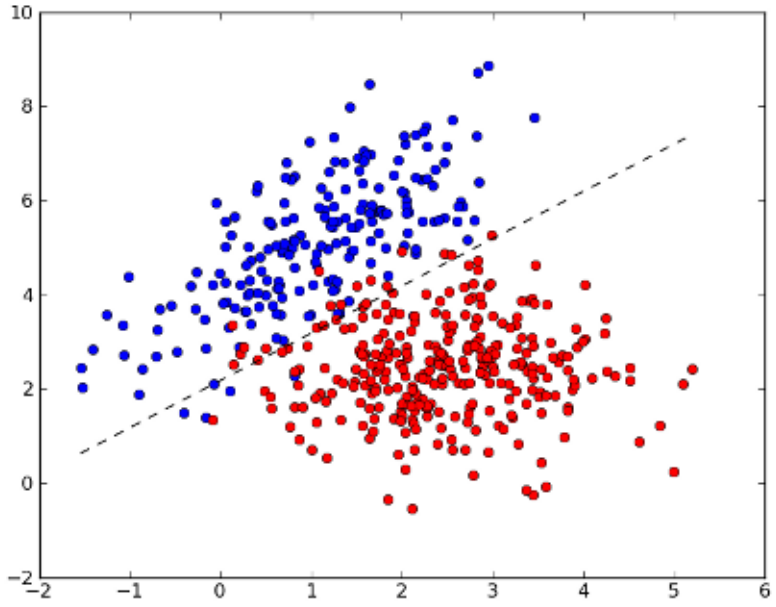

Figure 3.2: Basic Linear Regression

The coefficients θ are derived using the cost function, $J(\theta)$ which is shown in following equation. When the cost function is minimized the values of the coefficients are considered derived.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

Here,

m is the number of features

Y_i is an element of output set $Y = y_1, y_2, y_3, \ldots, y_n$ of the training data set.

3.7 Logistic Regression

When the dependent variable is categorical, then logistic regression is one of the most appropriate regression model. Logistic regression is a binary classifier, but multiple classification is possible using one vs. all model. If a data set D = (x,y) : x ϵ X , y ϵ Y where X is the independent variables and Y is the set of category. The hypothesis function h θ(x) is given below which is a sigmoid function,

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T X}}$$

Using this hypothesis function a curved line is drawn which is used to classify the data. To derive the coefficients a cost function J(θ) is used. The value of J(θ) is shown in following equation,

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log h_\theta(x_i)+(1-y_i)\log(1 - h_\theta(x_i))]$$

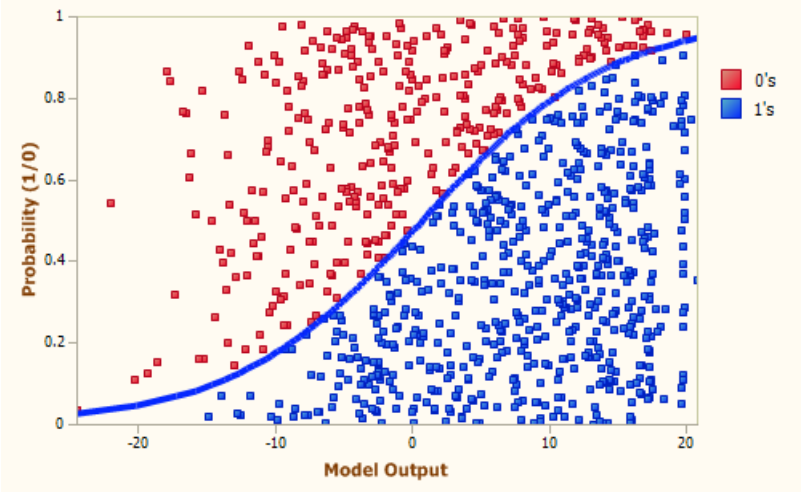

Figure 3.3: Basic Logistic Regression

After minimizing the cost function, the values of the coefficients are derived. The following equation is used to update the values of θ,

$$\theta_j = \theta_j - \frac{1}{m}\sum_{i=1}^{m}(h_j(x_i) - y_i)x_j$$

3.8 Artificial Neural Network

Inspired by the functionality of neural network, researchers tried make something similar and that is artificial neural network (ANN). The neurons takes some information as input via Dendrites, processes it and then sends to next neuron until it reaches destination. Similar to this ANN is shown in figure 3.4,

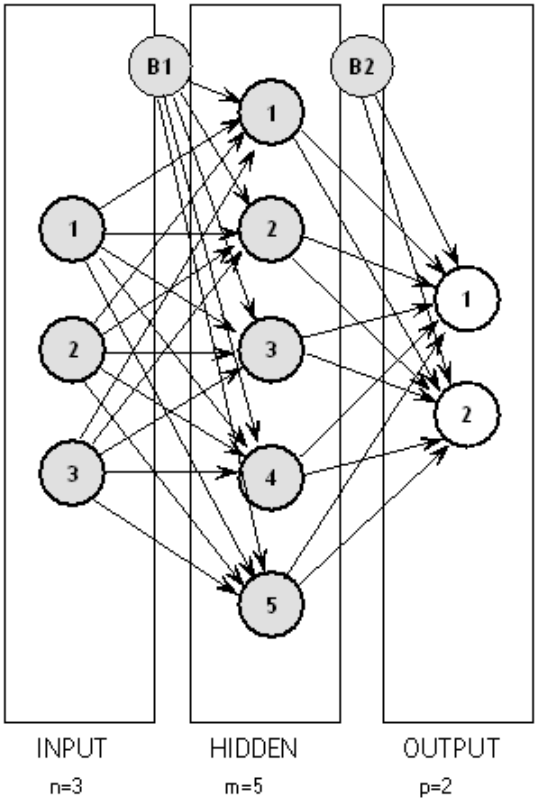

Figure 3.4: Basic Artificial Neural Network

Ann consists of few parts,

- Nodes which are shown as input, hidden and output layer.
- Edges which contains weights.
- Activation function. It defines whether a neuron will fire or not with a certain limitation.
- Bias input node that helps to find the solution faster and in other field as well. Except output

layer, all layers contains bias node.

As shown in the figure, the input node is connected with all node of hidden layer except bias. The hidden layers are connected with all output nodes. The weights are initialized with random numbers. ANN works in two phases: forward phase and backward phase. The Input node takes the inputs, processes it by multiplying the weight corresponded to the edge and sends it to next layer. Similarly hidden layer takes the input that was given, processes it using an activation function, then similarly the output is multiplied with the corresponding weight and sends it to output layer. The output layer takes the inputs, processes it and gives a corresponding output. This is called forward phase. In backward phase the weights are updated according to the output. In forward phase the following equation is used as activation function:

$$g(x) = \frac{1}{1+e^{-\beta}x}$$

Here, β is a random number.

The calculations for in each neuron k in hidden layer is given below:

$$h_k = \sum_{n=0}^{L} x_i v_{ik}$$

$$a_k = g(h_k) = \frac{1}{1+exp(-\beta h_k)}$$

Here, L is the number of input nodes. The calculations for in each neuron j in output layer is given below:

$$h_j = \sum_j a_k w_{jk}$$

$$y_j = g(h_j) = \frac{1}{1+exp(-\beta h_j)}$$

Here, N is the number of nodes in hidden layers.

The following equations are used for updating the weights in output layer:

$$\delta_o(j) = (y_j - t_j) y_j (1 - y_j)$$

$$w_{jk} \Leftarrow w_{jk} - \eta\delta_o(j)a_k$$

Here η is the learning rate. The following equations are used for updating the weights in hidden layer:

$$\delta_h(k) = a_k(1-a_k)\sum_{j=1}^{N} w_{jk}\delta_o(j)$$

$$v_{ik} = v_{ik} - \eta\delta_h(k)x_i$$

ANN works well with large data set. So, if the data set is too small then it will not perform well. The output often depends on number of iteration and number of node in hidden layer.

Chapter 4

Data Set

4.1 Data Collection

In our approach, we first tried to collect users Facebook profile data through Facebook Graph API with features like count, group activity no. of pictures uploads, check-in, birthday, place live in, education, age etc. But due to data restriction in current API version we failed to collect users profile data with sufficient features. Then we turn back and change our data collection approach. We created a form a with features Sex, Age, Religion and his/her Facebook status and collected approximately 250 users data. Below figure is showing the ratio of Male Female, Religion, Age. We tried to equalize the ratio of male female and religion.

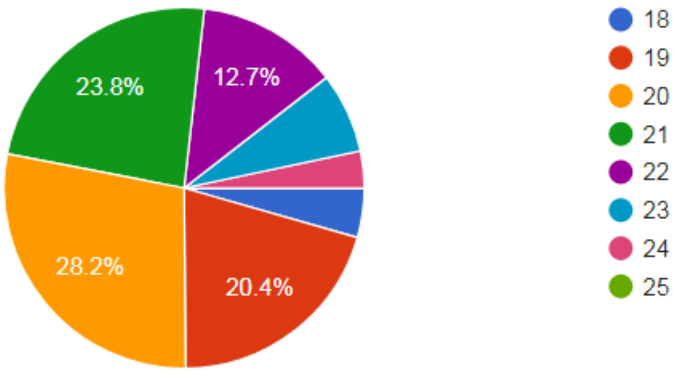

Figure 4.1: Users ratio according to Age

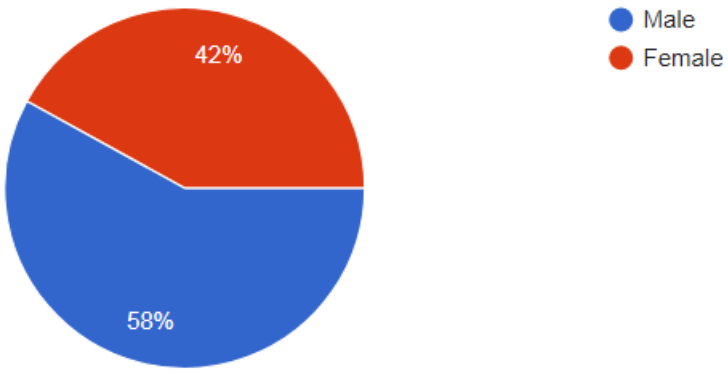

Figure 4.2: Users ratio according to Sex

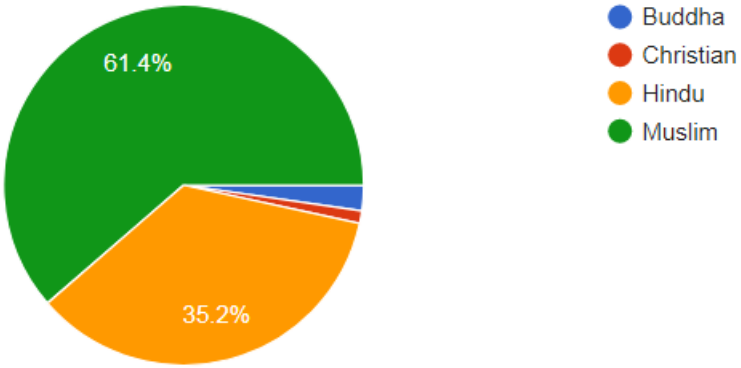

Figure 4.3: Users ratio according to Religion

4.2 Data Processing

In data formatting we tried to remove all kinds of garbage, English sentences, Facebook emoticon, Bangla numeric characters and other unusable symbols. We tried to keep significant Bangla words and remove non-significant Bangla words.

Chapter 5

Experimental Result and Analysis

In our experiment and result section we will discuss various classification algorithms and their accuracy on our data set individually and observe how they perform both individual features and joint features. Later we will compare those classification and select the best one for our user classification.

5.1 Experiment

5.1.1 Experiment with Gaussian Naive Bayes

In our approach we got significant performance for Sex and religion class but bad performance when we trained for both sex and religion. First of all, transformed our whole data set into TfidfVectorizer, used idf for that and remove all the unusual token form data set. Then we trained our model into four categories. First we trained with target class both sex and age which performed 58.18% accuracy which is not significant. Secondly, we trained for class age that performed better than previous class and got 76.36% accuracy.Thirdly, we trained for class religion that performed better than previous class and got 78.18% accuracy. Lastly, we got a significant performance for sex class with 94.54% accuracy which is better than all classifier model. The most important reason for this kind off accuracy is that sex class has only two features. That's why our Gaussian Naive Bayes model trained well for this class. Below Table-5.4 is showing the accuracy level for different classes.

Category	Accuracy (%)	Decision
Religion and Sex	58.18%	Not Good
Age	76.36%	Good
Religion	78.18%	Good
Sex	94.54%	Significant

Table 5.1: Accuracy of Gaussian Naive Bayes for different classes.

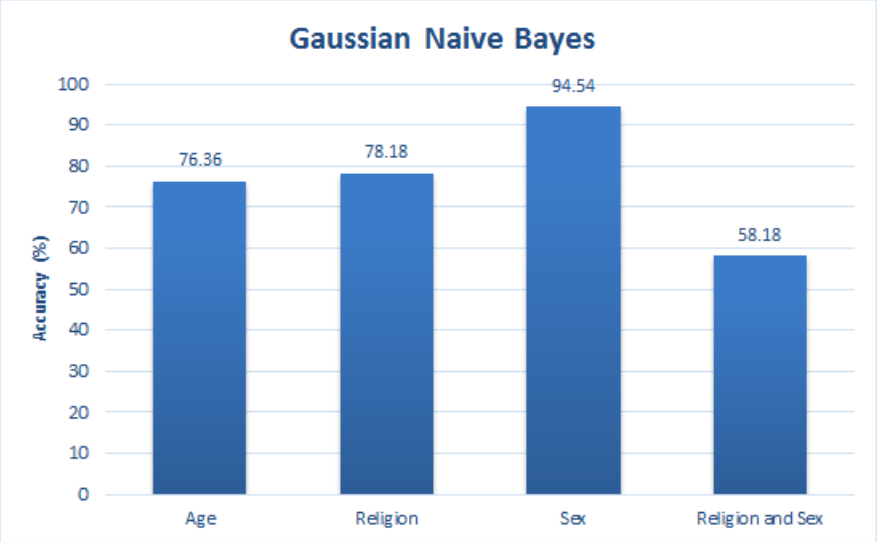

Figure 5.1: Performance diagram of Naive Bayes algorithm.

5.1.2 Experiment with SVM

As we stated before support vector machine draws a straight line between two class. That's why it always perform well in two class.First of all, we transform the whole data set into TfidfVectorizer. We set kernel='linear', set idf, penalty parameter C=1 and kernel coefficient gamma=1. We trained our data set for age and got result with performance 25% accuracy which was very low. Because the age class has eight features that are 18,19,20,21,22,23,24,25. We also trained our data set with SVM for religion class. Our religion class has 4 features with ratio that are Muslim 61.4% , Hindu 35.2% , Christian 1.1% and Buddist 2.3%. We got performance 76.36% accuracy. We tried our best to keep our data set in equal ratio. But due to lack of sufficient user we were unable to do that. In the other we trained our data set with sex class that has only two features that are Male or Female. Our model perform very well with 87.5% accuracy. We can say that this accuracy is

significant. We also trained our model for both Sex and Age that performed with 34.61% accuracy which is not significant. Table-5.3 shows all accuracy in percentage.

Category	Accuracy (%)	Decision
Age	25%	Not good
Sex and Age	34.61%	Not good
Religion	76.36%	Good
Sex	87.5%	Significant

Table 5.2: Accuracy of SVM.

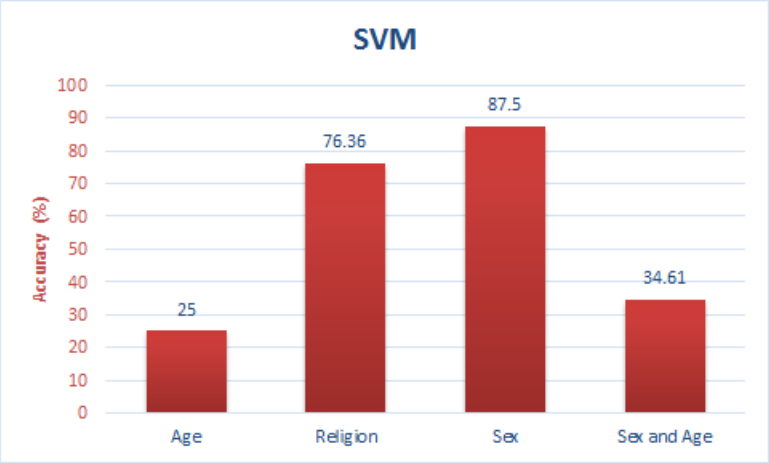

Figure 5.2: Performance diagram with SVM

5.1.3 Experiment with Linear Regression and Result

We applied linear regression on our data set and we got result 100% accuracy which was inaccurate. From analysis we came to know that our analysis was not good with linear regression. The major problem was data set. The data set is so big the better this algorithm perform well. We applied linear regression for classes age, sex and religion and in every case it gave performance with 100% accuracy. The below table shown the result of linear regression:

Category	Performance (%)	Decision
Age	100%	Overfitted
Sex	100%	Overfitted
Religion	100%	Overfitted

Table 5.3: Accuracy of linear regression.

5.1.4 Experiment with Logistic Regression

As we described before linear regression is not good for our analysis, in the same time we also got similar performance for logistic regression. Because of our small amount of data set we can get a better result for logistic regression. When we applied this model, it performs with accuracy 100% which we can say it is biased.

Category	Performance (%)	Decision
Age	100%	Overfitted
Sex	100%	Overfitted
Religion	100%	Overfitted

Table 5.4: Accuracy of logistic regression.

Category	Accuracy (%)	Decision
Age	100%	Overfitted
Sex	100%	Overfitted
Religion	100%	Overfitted

Table 5.5: Accuracy of Neural Network.

5.1.5 Experiment with Neural Network

Neural Network consists of input node, output node, activation function, hidden layer default 1, each layer connected with edge weight. Weight is updated in each iteration. In our model we used 1 hidden layer with 3000 hidden unit. Train our neural network for 200 iterations. As we stated before to use neural network with best performance we need a huge amount of data set. That's why in every case of our training we got accuracy with 100% which we can easily say that our model is biased. We also trained our neural network for different activation functions like relu, identity,

logistic and tanh. In every case we got bad performance and we took decision that our model is biased.

5.2 Comparison and Analysis

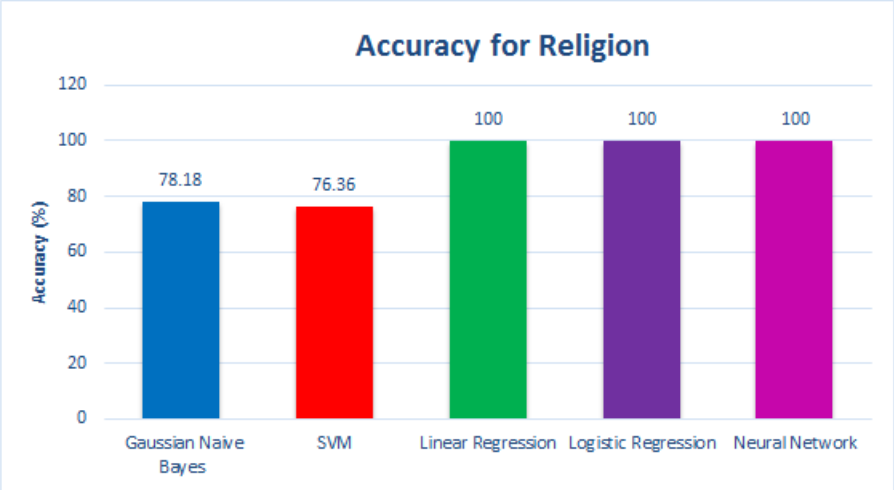

Figure 5.3: Performance diagram of all classification algorithm on Religion.

From the above diagram we can see that Gaussian Naive Bayes perform well for age, sex, and religion. But it performs very well for sex which is 94.54% although our data set is small So, choose Gaussian Naive Bayes classifier for sex classification with Facebook snippets. SVM also perform well for Sex and religion. And showed a bad performance for age classification. For religion and sex classification it performs 76.36% and 87.5% accuracy respectively. So, we can make decision that sex classification can be classified by Gaussian Naive Bayes model as it performs better than SVM.

On the other hand, we could not make decision for Artificial Neural Network, Linear Regression and Logistic Regression although they perform bad in training session. Because size of our data set was not enough big for train those algorithm.

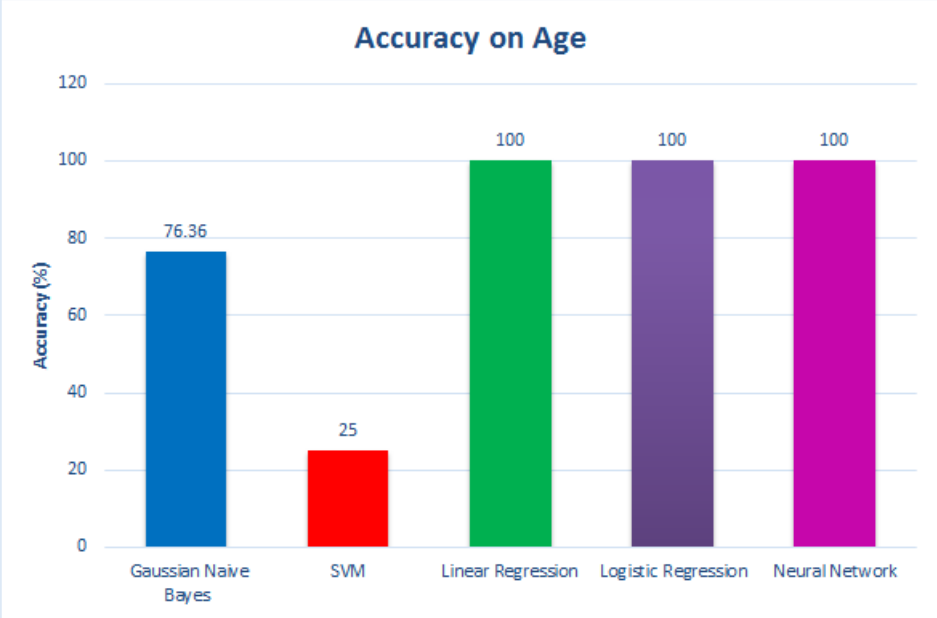

Figure 5.4: Performance diagram of all classification algorithm on Age.

5.3 Discussion

One important thing to note that we know there is no any machine that perform 100% accuracy. And we worked on users Facebook snippets with small data set and we got performance below 100% on other algorithm like Gaussian Naive Bayes and Support Vector Machine. So there is a lot of difference on performance between these two algorithm and Artificial Neural Network. We can not use Artificial Neural Network on this small data set. That's why we got over fit. Same case happen to Logistic Regression and Linear Regression. If we want to get better performance, we not only need large data set but also require same ratio of Religious, Sex and Age data. Otherwise, our model will be overfitted. Finally, we decide that we will use Gaussian Naive Bayes for Age, Sex and Religion classification for our small data set.

Chapter 6

Conclusion

6.1 Future Scopes

- The data set needs to be enlarged. ANN, Logistic Regression, Linear Regression etc. algorithms performs well with large data set.
- The age range of the data set is also to be enlarged. Wide range of age will help to divide the ages into age range which should perform better than just using a single age as people within certain age range performs similar activity.
- We can also consider other features for example emoticon, status timestamp, feeling of status etc.
- We can also focus on other social site for example LinkedIn as they are collecting many data from users.

References

[1] B. F. Marko Tkalcic and M. Schedl. *Personality traits and the relationship with (non-) disclosure behavior on Facebook*. Proceedings of the 25th International Conference Companion on World Wide Web, pp. 565â568, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2890085

[2] L. S. Alan Menk. *Predicting the human curiosity from users profiles on Facebook*. Proceedings of the 4th Spanish Conference on Information Retrieval Article No. 13, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2934743

[3] T. G. P. K. Yoram Bachrach, Michal Kosinski and D. Stillwell. *Personality and patterns of Facebook usage*. Proceedings of the 4th Annual ACM Web Science Conference, pp. 24â32, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2380722

[4] A. C. Reza Farahbakhsh, Xiao Han and N. Crespi. *Analysis of publicly disclosed information in Facebook profiles*. Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference, 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6785779/

[5] S. P. Thomas Theodoridis and Y. Kompatsiaris. *Assessing the reliability of Facebook user profiling*. Proceedings of the 24th International Conference on World Wide Web, pp. 129â130, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2742728

[6] Zhenhui Li, Ding Zhou, Yun-Fang Juan and Jiawei Han. *Keyword Extraction for Social Snippets*. Proceedings of the 19th international conference on World wide web, pp. 1143-1144 , 2010. [Online]. Available:
http://dl.acm.org/citation.cfm?id=1772845

[7] Aparna Garimella and Rada Mihalcea. *Zooming in on Gender Differences in Social Media*. 2016. [Online]. Available:
https://www.semanticscholar.org/paper/Zooming-in-on-Gender-Differences-in-Social-Media

[8] Jennifer Golbeck, Cristina Robles and Karen Turner. *Predicting Personality with Social Media*. CHI '11 Extended Abstracts on Human Factors in Computing Systems, pp. 253-262, 2011. [Online]. Available:
https://www.semanticscholar.org/paper/Zooming-in-on-Gender-Differences-in-Social-Media

[9] Yin Zhu, Erheng Zhong, Sinno Jialin Pan, Xiao Wang, Minzhe Zhou and Jiawei Han. *Predicting user activity level in social networks*. Proceedings of the 22nd ACM international conference on Information Knowledge Management, pp. 159-16, 2013. [Online]. Available:
http://dl.acm.org/citation.cfm?id=2505518

[10] Jalal S. Alowibdi, Ugo A. Buy and Philip Yu. *Language Independent Gender Classification on Twitter*. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 739-743, 2013. [Online]. Available:
http://dl.acm.org/citation.cfm?id=2492632CFID=983309617CFTOKEN=98982289