

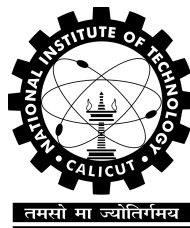
Classification of Age groups in Social network

*Report of the project submitted in partial fulfillment of
the requirements for the award of the degree of*

**Bachelor of Technology
in
Computer Science and Engineering**

Submitted by
Boggarapu Haranath (B140867CS)
Gugloth Mallesh (B140922CS)
P Santhosh Kumar (B140814CS)

Under the guidance of
Sreenu Naik Bhukya



Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
Calicut, Kerala, India – 673 601

May 1st, 2018

Acknowledgements

This project is a culmination of hard work and efforts made by many people, providing support and encouragement to us. We express our sincere gratitude to everyone who helped us in completing this undertaken task successfully.

We thank our project guide Sreenu Naik Bhukya, Asst. Prof., NIT Calicut for his guidance, support, motivation and facilities provided.

We thank all teaching and non teaching staff, for their cooperation, discussions and suggestions.

B Haranath
G Mallesh
P Santhosh kumar

May 1, 2018
National Institute of Technology Calicut

DECLARATION

"I hereby declare that this submission is my own work, and that to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text".

Place:

Date:

Signature:

Name: Boggarapu Haranath
Roll No.: B140867CS

Name: Gugloth Mallesh
Roll No.: B140922CS

Name: P Santhosh Kumar
Roll No.: B140814CS

CERTIFICATE

*This is to certify that the project report entitled: **Classification of Age groups in Social network** submitted by **Boggarapu Haranath (B140867CS), Gugloth Mallesh (B140922CS), P Santhosh Kumar (B140814CS)** to National Institute of Technology Calicut towards partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide record of the work carried out by him/her under my supervision and guidance.*

Sreenu Naik Bhukya
(Project Guide)

Dr. Saidalavi Kalady
(Head of the Department)

Place:

Date:

Office Seal

Abstract

There is big amount of records in social networks, wherein human beings post their opinion on a topic, or proportion their records. but humans frequently dont provide their private records, like gender, age and different demographics.

furthermore, with modern technology it is quite easy to create profile with fake age, gender and place which provides criminals an clean way to mislead. there are numerous extra other packages like sentiment analysis and etc as a result we can analyze the text messages published by using the person on social community platform.

As in keeping with the research performed to date, age is one of the crucial parameter in the consumer profile which exhibits the critical records approximately the everyday behavior among same age group users. An analysis is executed with dataset which includes relevant parameters are taken into consideration for the category. in this challenge we classified the users into two age organizations teens and adults the usage of special device learning algorithms. among these kinds of algorithms DCNN stands out to be the great amongst all of them attaining the high-quality performance with an accuracy of 94%.[1]

Contents

Acknowledgements	1
1 Introduction	1
1.1 Background and Recent Research	1
1.1.1 Observations	1
1.1.2 Literature Survey	2
1.2 Motivation	3
2 Problem Definition	4
3 Work Done	5
3.1 Data Collection	5
3.2 Data Processing	5
3.3 Design	5
3.3.1 Convolution	6
3.3.2 Max pooling/Sub-Sampling	7
3.3.3 Activation	7
3.3.4 Fully connected	8
3.4 Feature extraction	8
3.5 Implementation	9
3.5.1 Local receptive fields	9
3.5.2 Shared weights	9
3.5.3 Pooling or Sub sampling	10
3.5.4 Tools/hardware/Language used for development: . . .	10
4 Future Work	12
5 Conclusion	13
References	14

List of Figures

3.1	Flow chart for building a model	6
3.2	Testing	6
3.3	Pooling[8]	7
3.4	Convolution Neural Networks[6]	9
3.5	Data extraction[9]	10

Chapter 1

Introduction

1.1 Background and Recent Research

1.1.1 Observations

With regular use of the internet, these days customers spend hours surfing on social networks about numerous topics. these sports may be analyzed to assess customer satisfaction that may be a very useful facts for provider providers and product providers.[10]

presently, there's a difficulty and a greateffort to analyze facts from online so-cial networks to predict facts thatmay re ect exceptional components of the cur-hire fact. however, the casual andshort sentences with many versions oflanguage do important the have a look at of someparameters to improve the records analysis.amongst them is the age thatcan at once in uence the final sen- ti-ment of a sentence. not unusual individual-istics, observed at some point of every segment of existence,are taken into account in this kind ofanalysis; especially, those characteris-tics are without a doubt one of a kind in the teenagerand grownup age agencies. it's far vital tonote that in some social networks, theuser age isn't to be had both by using the so-cial community itself or even through the userfor discretion reasons; consequently,the dedication of a method to predicthe users age is relevant in the sentiment-analysis.[4]2 hassle StatementTo determine the characteristics ofteenager and adults age organizations, consid-ering the writing style and each customer-records and profile. one of the maximum rel-evant parameter contained in the userprofile is the age institution showing thatthere are typical behaviors among theusers of identical age organization.

To display parameters used in this re-seek can attain a excessive accuracy

for determining age businesses of facebook users. Validate the usefulness of the proposed model for classifying age organizations. As it was proven that on fb, informing the age within the profile description isn't a common addition and consequently our studies would now not provide reliable consequences. The aim of this task is therefore to create a model that would expect the age group of user with the assist of person profile information. The version is applied using artificial neural networks.[2]

1.1.2 Literature Survey

There is a huge body of extremely good research on improving social statistics with demographic attributes. Some of the most exciting paintings on detecting age group from social statistics has been in the computer vision wherein age is expected using the user pictures.[3]

There has been research on prediction of age by means of lexical analysis of the messages posted on social networks. Gender best friend older human beings react definitely and use fewer words. It's pretty obvious from previous studies that age prediction works well for the more youthful people than older people. Research on dating between age institution and traits of writing appears to show better result in comparing grownup and teenagers.[4]

Inside the early level of research on predicting the age of users, consumer name also been used a source of demographic information for predicting the age of users in social community. Chance distributions of birth years given age is generated which display that for some names distribution become sharply peaked. The age records isn't constantly seasoned-vided in some social networks, for instances, fb. After verifying that this records should certainly modify the results of several analysis, a few research have worked on trying to expect it. But, it became established that on facebook, informing the age in the profile description is now not a not unusual habit and therefore those research would not provide reliable consequences.[5]

1.2 Motivation

With the advancement in Artificial Intelligence ,the ability to learn for computers has enormously increased .This helps in reducing lot of work,money and time for the mankind .As there are lot of applications that can be created with the help of Deep learning and one of them is age group classification ,where many of the e-commerce websites uses this algorithm to know which age group of people like the products they want to buy. This motivates me to make a similar algorithm which predicts the age group based on their data provided on their social networking sites.

Chapter 2

Problem Definition

To determine the characteristics of teenager and adults age groups, considering the writing style and both users history and profile.

To show parameters used in this research can reach a high accuracy for determining age groups of twitter users and Validate the usefulness of the proposed model for classifying age groups.

The aim of this project is hence to create a model that could predict the age group of user with the help of user profile information. The model is implemented using artificial neural networks.

Chapter 3

Work Done

3.1 Data Collection

To extract labels we crawled the Facebook Graph and downloaded the user profiles and news feed. To do this we implemented a crawler that using Facebook access tokens of different users both young and adults. Our crawl downloaded about more than 3000 users profile description fields. We downloaded name, recent news feed, comments, user follows, user followers for each id.

3.2 Data Processing

The dataset comprises of user profile information such as name, education, friends, likes, news feed, education, date of birth. Some of the fields like date of birth, number of friends, likes may not contain information, so we need to fill NULL values with the average values of the fields. Punctuation and emojis in the news feed are need to be identified in the message posted by the user.[7] Dataset contains story which tells the post contains URL link or photos are added.

3.3 Design

We implemented these in four stages:

To train the Neural Network, if we are given the values of X and Y for various examples in training set, hidden layers will adjust automatically. Every Hidden Unit is connected by all input units, densely connected be-

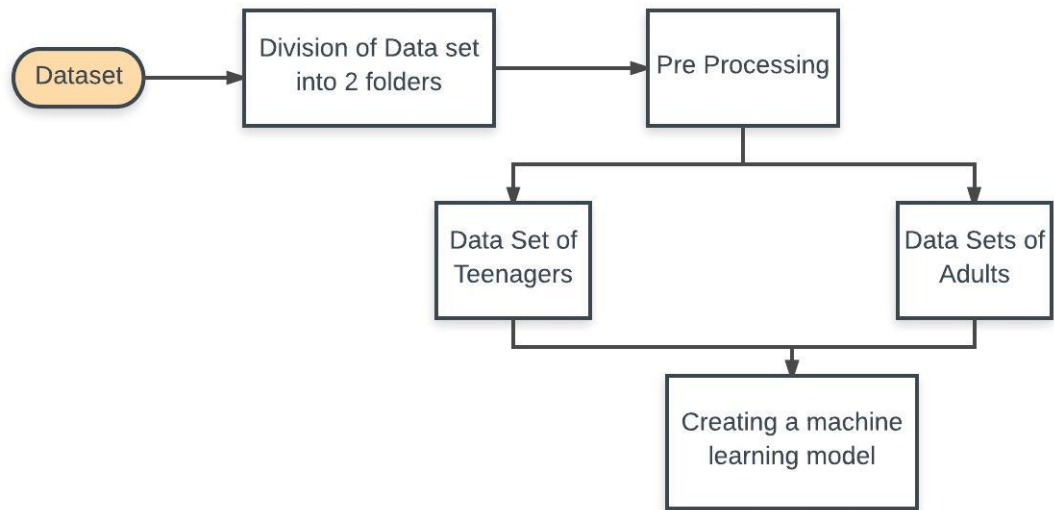


Figure 3.1: Flow chart for building a model

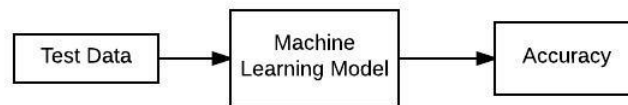


Figure 3.2: Testing

tween Input layer and Hidden layer.

There are four main steps in CNN convolution, subsampling, activation and full connectedness. The most popular implementation of the CNN is the LeNet, after Yann LeCun. The 4 key layers of a CNN are Convolution, Sub-sampling, Activation and Fully Connected.

3.3.1 Convolution

The first layers that receive an input signal are known as convolution filters. Convolution is a manner wherein the network tries to label the entire signal by means of referring to what it has found out in the past. If the input signal appears like a previous cat image while it has visible earlier than, the cat reference signal will be mixed into, or convolved with, the entire signal. The ensuing output-positioned signal is then exceeded directly to the next layer. Convolution has the excellent assets of being translational invariant. Intu-

itively, this manner that every convolution filter represents a feature of interest (e.g. whiskers, fur), and the CNN algorithm learns which capabilities contain the resulting reference. The outputted signal strength isn't depending on wherein the features are placed, but honestly whether the functions are gift. therefore, a cat could be sitting in exclusive positions, and the CNN algorithm would nevertheless be able to recognize it.

3.3.2 Max pooling/Sub-Sampling

Inputs from the convolution layer can be smoothened to lessen the sensitivity of the filters to noise and variations. This smoothing manner is called subsampling, and may be completed by means of taking averages or taking the most over a sample of the sign. Examples of subsampling strategies (for image signals) consist of reducing the dimensions of the photo, or reducing the colour contrast throughout crimson, green, blue (RGB) channels.

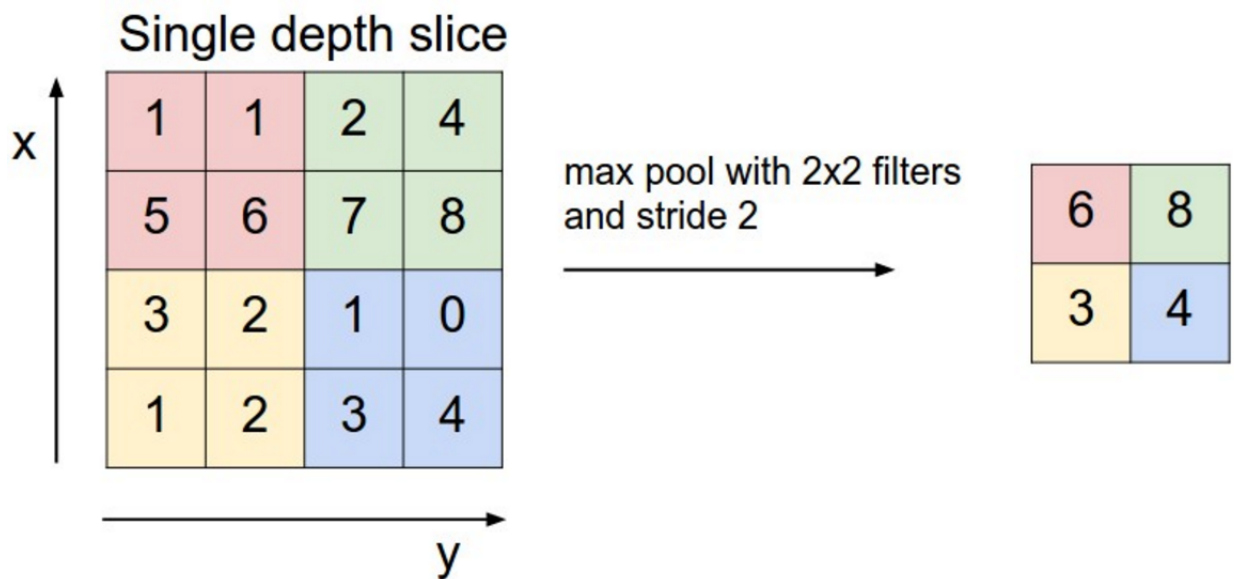


Figure 3.3: Pooling[8]

3.3.3 Activation

The activation layer controls how the signal flows from one layer to the following, emulating how neurons are fired in our mind. Output indicators which are

strongly associated with beyond references might activate extra neurons, enabling indicators to be propagated extra efficaciously for identification. CNN is well suited with a extensive variety of complicated activation capabilities to model sign propagation, the maximum common function being the Rectified Linear Unit (ReLU), which is favored for its faster training velocity. The summation of inputs $w_1 * x_1$, $w_2 * x_2$, $w_3 * x_3$are passed as a parameter to activation function $f(x)$ as $f(\sum w * x)$

f is called the activation function. Here, we take sigmoid function as Activation function.

$$f(z) = 1 / (1 + \exp(-z))$$

thus, our single neuron corresponds precisely to the enter-output mapping described with the aid of logistic regression. even though these is find of sigmoid feature, it is worth noting that some other commonplace preference for f is the hyperbolic tangent, or tanh, function.

$$f(z) = \tanh(z)$$

3.3.4 Fully connected

The last layers in the network are fully connected, meaning that neurons of preceding layers are connected to every neuron in subsequent layers. This mimics high level reasoning where all possible pathways from the input to output are considered.

Here S1, S2 are Max Pooling/ SubSampling layers and C1, C3 are Convolution Layers and the last layer is Fully Connected layer. This is the layer pattern we are following.

3.4 Feature extraction

We choosed certain parameters for predicting age group. Since, Convnets (Convolutional Neural Networks only need structured data (like 1D arrays or 2D arrays), we choose numerical parameters. Our parameters are[9]

- Message Length
- Number of Posts
- Total Post likes
- Total Comments
- Number of Friends
- Number of page Likes
- Number of Emoji's
- Number of Hashtag's

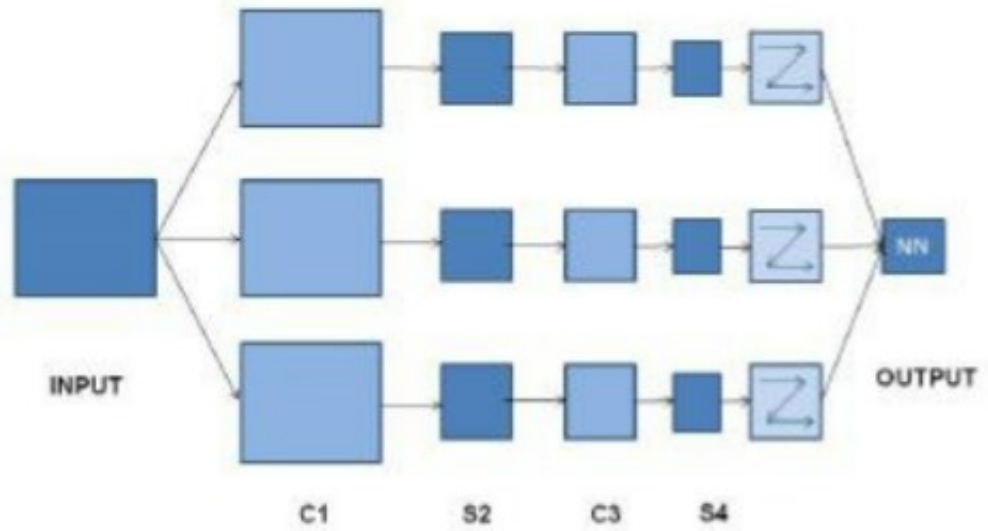


Figure 3.4: Convolution Neural Networks[6]

3.5 Implementation

CNN also known as covnets are defined on the basis of following principles in our project.

3.5.1 Local receptive fields

A CNN does not have a receptive field. A neuron, or a kernel, or a convolution layer has a receptive field, which is the size of the kernel. Typically 3x3.

3.5.2 Shared weights

Shared weights basically means that the same weights are used for two layers in the model to take the advantage of Parameter reduction.

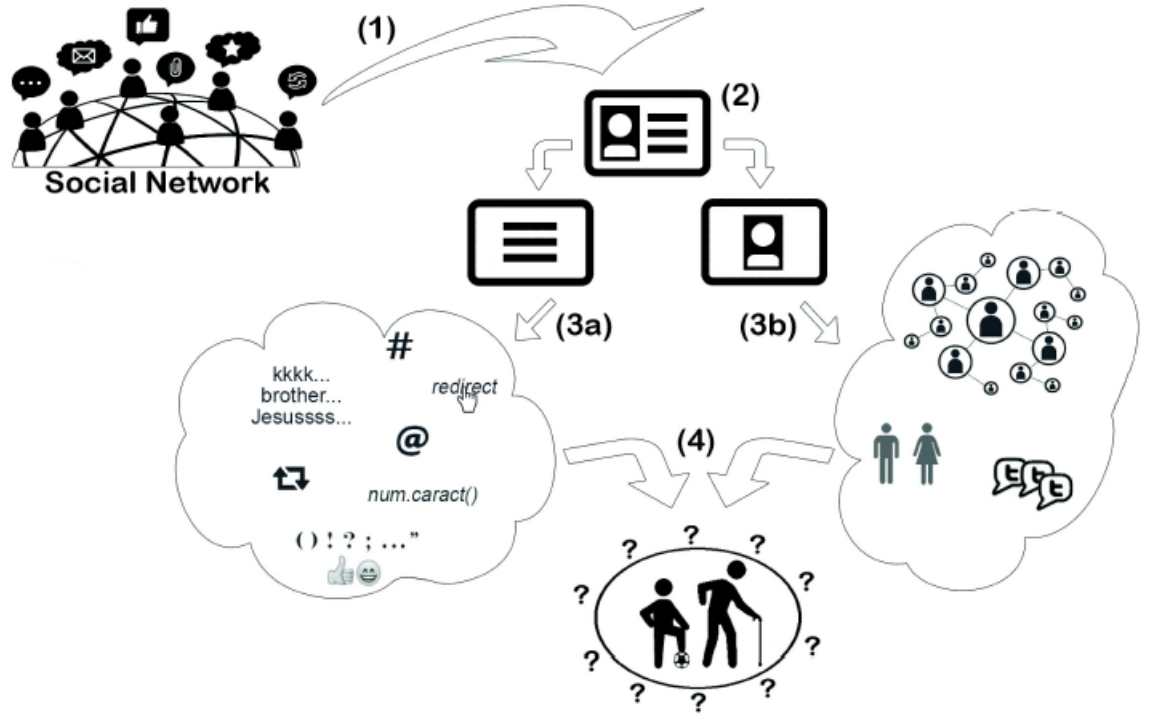


Figure 3.5: Data extraction[9]

3.5.3 Pooling or Sub sampling

In-puts from the convolution layer can be smoothed to reduce the sensitivity of the filters to noise and variations. This smoothing process is called subsampling, and can be achieved by taking averages or taking the maximum over a sample of the signal. Examples of subsampling methods (for image-signals) include reducing the size of the image, or reducing the color contrast across red, green, blue (RGB) channels.

3.5.4 Tools/hardware/Language used for development:

Anaconda:

Anaconda is a Python IDE. It provides a large selection of packages and commercial support. It is an environment manager, which provides the facility to create different python environments, each with their own settings. Conda, the Anaconda's own package manager, is used for updating, installing and

executing ana-conda and its bunch of packages like numpy, scipy, ipython notebook etc It helps in switching between environments in our local machine.

Tensorflow

TensorFlow is an machine learning library, which can be used for high-level implementation of various ML algorithms in Python. However, it is used primarily for deep learning, Because most deep learning uses GPUs, it is common to use deep learning frameworks that implement common operations in GPUs (having more computational units than CPU, Deep Neural Networks (DNN) are structured in a very uniform manner such that at each layer of the network thousands of identical artificial neurons perform the same computation. Therefore the structure of a DNN fits quite well with the kinds of computation that a GPU can efficiently perform).

Keras:

Keras is an open source neural machine learning library written in Python.

Chapter 4

Future Work

The Proposed models gave satisfactory results, as there is no model that gives 100% accuracy there is still a scope of improvement in the area. One of the interesting issues to learn in the future would be developing a model which also uses the images of the user along with profile information and posts.

We used the number of friends as parameter but we can also number of followers but this information cannot be extracted using facebook graph API. The results which are obtained in this project can be used in sentiment analysis to enhance the performance.

Chapter 5

Conclusion

At present, we are dealing with 1-D arrays. But already, Deep Learning (Convolutional Neural Networks) is dealing with images and showed an enormous progress in the field of image recognition field. Hence, it is more effective to implement Convnets on images rather than 1-D arrays. Data scientists are trying all possible chances to convert every unstructured data to structured data to achieve high progress since Convnets gives best performance for structured data.

References

- [1] https://www.researchgate.net/publication/317129260_Age_Groups_Classification_in_Social_Network_Using_Deep_Learning
- [2] J. Huang, J. Lu and C. X. Ling. Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy,, Proceedings of Third IEEE ICDM, 19-22 November 2003, pp. 553-556. doi:10.1109/ICDM.2003.1250975.
- [3] P. S. Vadivu and V. K. David. "Enhancing and Deriving Actionable Knowledge from Decision Trees,, IJCSIS, Vol. 8, No. 9, 2010, pp. 230-236
- [4] J. van de Loo, G. De Pauw, and W. Daelemans "*Text-based age and gender prediction for online safety monitoring*", Comput. Linguistics Netherlands, vol. 5, no. 1, pp.46 60, Dec. 2016
- [5] Antonio A. Morgan- Lopez, Annice E. Kim, Robert F. Chew, Paul Ruddie. "Predicting age groups of Twitter users based on language and meta- data features," ,Available at: <https://doi.org/10.1371/journal.pone.0183537> 4
- [6] <https://ieeexplore.ieee.org/mediastore/IEEE/content/media/6287639/7859429/7932459/guima1-2706674-glance.gif>
- [7] J. A. B. L. Filho, R. Pasti, L. N. de Castro. "*Gender classification of twitter data based on textual meta-attributes extraction*" Adv. Intell. Syst.Comput., vol. 444, pp. 1025- 1034, March 2016.
- [8] <https://qph.ec.quoracdn.net/main-qimg-a3f20fafa7e08e3bf5ba3a99e696c46b-c>
- [9] <https://ieeexplore.ieee.org/mediastore/IEEE/content/media/6287639/7859429/7932459/guima2-2706674-glance.gif>
- [10] Using deep learning age groups classification in social networks <https://ieeexplore.ieee.org/document/7932459/>