Classification of Age Groups in Social Network

CS4090

Final Report

B.Haranath, G.Mallesh, P.Santhosh Kumar
Guided By: Sreenu Naik Bhukya

May 1, 2018

Abstract

Predicting accurate demographic information about the users of information systems is a problem of interest in personalized search, ad targeting, and other related fields. Despite such broad applications, most existing work only considers age prediction as one of classification, typically into only a few broad categories. Here, we consider the problem of age classification in social networks.

1 Introduction

With constant use of the Internet, these days users spend hours browsing on social networks about diverse topics. These activities can be analyzed to assess customer satisfaction that is a very useful information for service providers and product suppliers.

Currently, there is a concern and a great effort to analyze data from online social networks to predict information that may reflect different aspects of the current reality. However, the informal and short sentences with many variations of language do necessary the study of some parameters to improve the data analysis. Among them is the "age" that can directly influence the final sen- timent of a sentence. Common characteristics, found during each phase of life, are taken into account in this type of analysis; specifically, those characteristics are clearly different in the teenager and adult age groups. It is important to note that in some social networks, the user age is not available either by the social network itself or even by the user for discretion reasons; as a consequence, the determination of a method to predict the users age is relevant in the sentiment analysis.[4]

2 Problem Statement

To determine the characteristics of teenager and adults age groups, considering the writing style and both users' history and profile. One of the most relevant parameter contained in the user profile is the age group showing that there are typical behaviors among the users of same age group.

To show parameters used in this research can reach a high accuracy for determining age groups of facebook users. Validate the usefulness of the proposed model for classifying age groups. As it was verified that on facebook, informing the age in the profile description is not a common habit and therefore our studies would not provide reliable results. 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.

3 Literature Survey

There is a large body of excellent research on enhancing social data with demographic attributes. Some of the most exciting work on detecting age group from social data has been in the computer vision where age is predicted using the user images.[1]

There has been research on prediction of age by lexical analysis of the messages posted on social networks. Generally older people react positively and use fewer words. Its quite evident from previous research that age prediction works well for the younger people than older people. Research on relationship between age group and characteristics of writing seems to show better result in comparing adult and teenagers.

In the early stage of research on predicting the age of users, user name also been used a source of demographic information for predicting the age of users in social network. Probability distributions of birth years given age is generated which show that for some names distribution was sharply peaked.

The age information is not always provided in some social networks, for instances, Facebook. After verifying that this information could actually alter the results of several analysis, some research have worked on trying to predict it. However, it was verified that on facebook, informing the age in the profile description is not a common habit and therefore these studies would not provide reliable results.

4 Work Done

4.1 Data Collection

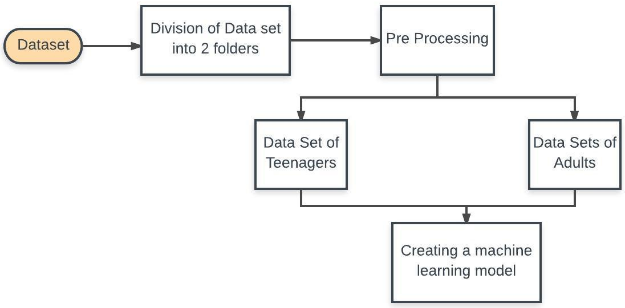
To extract labels we crawl the Facebook Graph and download the user profiles an news feed. To do this we implemented a crawler that using Facebook access tokens ok different users both young and adults. Our crawl downloaded more than 2000 users profile description fields. We downloaded name, recent news feed, com- ments, user follows, user followers for each id.[3]

4.2 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. Dataset contains story which tells the post contains URL link or photos are added.

4.3 Design

We implemented this in four stages:



Dataset
Division of Data set into 2 folders
Pre Processing
Data Set of Teenagers
Data Sets of Adults
Creating a machine learning model

Figure 1: Flow chart for building a model

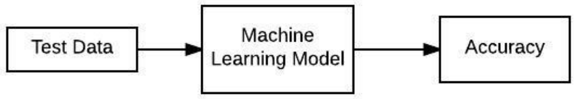


Figure 2: Testing

To train the Neural Network, if we 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, densily connected between 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, Subsampling, Activation and Fully Connected.

Convolution:

The first layers that receive an input signal are called convolution filters. Convolution is a process where the network tries to label the input signal by referring to what it has learned in the past. If the input signal looks like previous cat images it has seen before, the cat reference signal will be mixed into, or convolved with, the input signal. The resulting output signal is then passed on to the next layer.

Convolution has the nice property of being translational invariant. Intuitively, this means that each convolution filter represents a feature of interest (e.g whiskers, fur), and the CNN algorithm learns which features comprise the resulting reference. The out- put signal strength is not dependent on where the features are located, but simply whether the features are present. Hence, a cat could be sitting in different positions, and the CNN algorithm would still be able to recognize it.

Max pooling/Sub-Sampling: Inputs from the convolution layer can be smoothened 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.

Activation:

The activation layer controls how the signal ows from one layer to the next, emulating how neurons are fired in our brain. Output signals which are strongly associated with past references would activate more neurons, enabling signals to be propagated more efficiently for identification. CNN is compatible with a wide variety of complex activation functions to model signal propagation, the most com- mon function being the Rectified Linear Unit (ReLU), which is favored for its faster training speed. The summation of inputs w1*x1, w2*x2, w3*x3.....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 exactly to the input-output mapping defined by logistic regression. Although these is use of sigmoid function, it is worth noting that another common choice for f is the hyperbolic tangent, or tanh, function.

f(z) = tanh(z)

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.

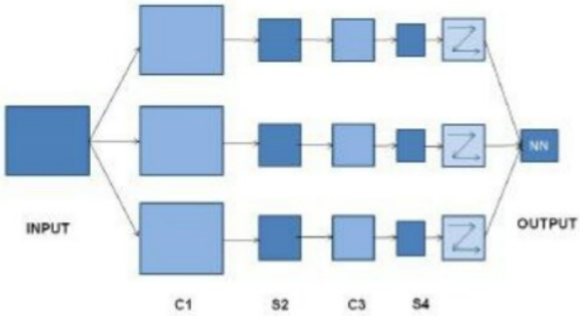


Figure 3: Convolutional Neural Network[2]

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

4.4 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

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

4.5 Implementation

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

Local Receprive 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.

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

Pooling or Sub Sampling: Inputs from the convolution layer can be smoothened 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 nels.

Tools/hardware/Language used for development:

Anaconda

Anaconda is a Python IDE. It provides 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 Anacondas own package manager, is used for updating, installing and executing anaconda and its bunch of packages like numpy, scipy, ipython notebook etc It helps in switching between environments in our local machine.

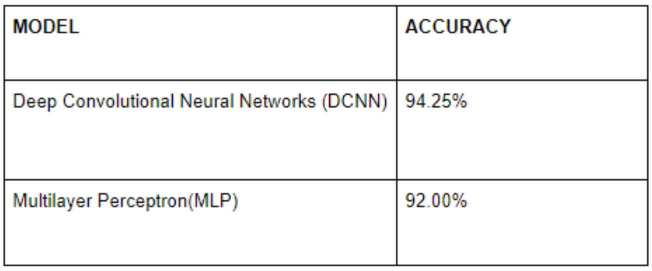
Tensorflow

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.

4.6 Results and Analysis



MODEL	ACCURACY
Deep Convolutional Neural Networks (DCNN)	94.25%
Multilayer Perceptron(MLP)	92.00%

Figure 4: Results

And the performance of DCNN in terms of Precision, Recall and F-measure is much better than MLP (Multi

Layer Perceptron) with values of DCNN (0.929, 0.936, 0.930) and values of MLP are (0.882, 0.869, 0.880).

5 Future Work and Conclusions

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 recognization 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 perfomance for structured data.

References

[1] 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

[2] *Age Groups Classification in Social Network Using DCNN*, *https://www.ieeeexplore.ws/document/7932459/.*, Volume:inclusive pages, date of publication.

[3] Antonio A. Morgan- Lopez, Annice E. Kim, Robert F. Chew, Paul Ruddle. "Predicting age groups of Twitter users based on language and meta- data features,",Available at: https://doi.org/10.1371/journal.pone.0183537 4

[4] J. A. B. L. Filho, R. Pasti, L. N. de Castro. *"Gender classiffication of twitter data based on textual meta-attributes extraction"*Adv. Intell. Syst.Comput., vol. 444, pp. 1025- 1034,March 2016.