An Optimal Deep Learning Approach for Classification of Age Groups in Social Network

Anil Kumar Swain
Department of Computer Sc. & Engineering, NIT Meghalaya, Shillong, India
anilkumarswain@nitm.ac.in

Bunil Kumar Balabantaray
Department of Computer Sc. & Engineering, NIT Meghalaya, Shillong, india
bunil@nitm.ac.in

Jitendra Kumar Rout
School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India
jitendra.routfcs@kiit.ac.in

Suneeta Satpathy
Department of Computer Sc. & Engineering, CEB, BPUT, Odisha, India
suneeta1912@gmail.com

ABSTRACT

There is huge amount of data in social networks, where people post their opinion on a topic, or share their information. But people often don't provide their personal data, like gender, age and other demographics. Research can be done on this data to develop applications of sentiment analysis, but the success rate is restricted by the number of words in the dictionaries as they do not consider all the words which reflect the sentiment in our messages as most of the communication on social networks is non-standard language with small messages. Moreover, with contemporary technology it is quite easy to create profile with false age, gender and location which provides criminals an easy way to deceive. Thus we can analyze the text messages posted by the user on social network platform. As per the research done so far, age is one of the important parameter in the user profile which reveals the important information about the typical behavior among same age group users. An analysis is done with more than 4000 tuples which contains relevant parameters like number of friends, length of message, number of likes, number of hash tags and comments are considered for the classification. In this study, we use the user profile information for the prediction of age group, which we collected using Facebook API. In this paper we classified the users into two age groups teenagers and adults using different Machine learning algorithms like deep convolutional neural networks, Multilayer perceptron, Random forest , SVM and Decision trees. Among all these algorithms deep convolutional neural network stands out to be the best among all of them reaching the best performance with an accuracy of 94%.

Keywords: Facebook, DCNN, Multilayer Perceptron, SVM, Decision trees.

1. INTRODUCTION

The Social networking platforms have now become a part of modern lifestyle with people spending majority of the time online. It has become medium for communication, and connecting with people across the world. Now a days, people have shown interest to express their opinions on various topics through posts on social networking sites like facebook. In order to evaluate the quality of services, it's important to measure the customer satisfaction so as to see if the actual delivered performance of service matches with the expectations. Therefore it is essential to know the sentiment of the customers regarding the services.

Research has already been done on age classification in twitter where the whole user group is divided into teenger group or an adult group using user demographics and user profile information. [1]

For social media monitoring and for a broader overview of the wider public opinion on varied topics sentiment analysis plays a very important role. Few of them are already implemented for instances, for capturing patient experience in hospitals [2], crime pattern detection using online social media [3], leveraging political campaigns from opinions expressed on social media[4],predicting price of stock from stock market indicators like Sensex and Nifty[5].

Facebook having more than 2.13 billion monthly active users provides access to posts and comments that can be collected and analyzed. These informal posts and comments have several variations with respect to slang, emoticons, length of post or a comment. In addition there may be several spelling errors or typographical errors.All these provide parameters for data analysis.

The variations in the linguistic characteristics can be analyzed and a person's age or gender can be predicted. There are few studies considering this aspect of sentiment analysis for instances, Anton Alekseev[6] investigated the problem of predicting information relating to demography based on texts generated by user from a large social network. Wiesław Wolny[7] examined how twitter data can be analyzed using emotion Several studies classified gender in social networks based on linguistic characteristics or patterns for instances, Qunazeng [8] researched on the problem of predicting gender based on images posted online, Jalal S. Alowibdi[9] investigated on gender classification in twitter from language independent parameters. Furthermore, several studies proved that age classification is much more challenging than gender classification in social networks.

For this purpose of age classification two phases of life are taken into consideration. These two phases of life, such as, teenagers and adults are clearly distinct.When an user is classified under a teenager or an adult, the user can get more personalized overview when it comes to recommendation systems in real time environment thus enhancing user experience in social networking sites and also e- commerce sites.

This research has been performed using facebook but it can be extended to other social networks because the parameters will remain the same.

In this research, several characteristics of a particular facebook user are considered as parameters, such as, length of message, number of posts, total post likes, total number of comments, number of friends, number of page likes, number of emoji's and the number of hashtags to classify the teenager and adult age group. All these parameters were found to accurately classify the age groups into two labels. Furthermore, several machine learning algorithms such as Multi perceptron, Random forest were tested and Deep Convolutional Neural Network was found to exhibit higher performance.

In this paper. Section II presents some related works on age groups classification and how machine learning is used in classifying teenagers and adults. In Section III the proposed model is explained. The results are discussed in Section IV and then some discussions are explained in Section V. Finally, the conclusion and future work is in Section VI.

2. RELATED WORK

TIn this section, we study the relevant parameters that impact the classification of age groups. Some recent studies prove that age parameter enhances the performance of sentiment analysis. Machine learning algorithms used for this classification problem are also discussed.

Characteristics of age groups

Persons in range between 13 and 21 are considered as teenagers. Behavior of different age groups is quite evident from their writing style or way of expressing their views on a topic. Teenagers' post or share lot of information on social networking platforms like facebook.

Many people in social networks like facebook do not provide age information. But Age is one of the important demographic which is used for research or to improve the results of several surveys or analysis. Research is also done to predict the age and categorize people on social media to use it in different applications like online marketing etc. There are different ways to predict age group, one of them is based on linguistic differences. Writing style, punctuation and speech pattern varies among different age groups.

As it is not mandatory to mention the age information on facebook many of the users skip that field for privacy issues. One strategy used is to search timeline of an user and filter the posts which give information about his birthday, work experience. For example, joined in XYZ organization tells that user is an adult. But these strategies would not produce better results.

Topics that teenagers prefer to discuss on social media are quite different from adults. Teenagers show their interest in topics that occur in day to day life. Posts on school, college and friends are more frequent in this age group.

Adults tend to be more careful when they post on facebook. They don't write much in their posts but sentences reflect positivity more than negation with less punctuation.
Teenagers post lot of information on social networking platform like facebook whereas adult do not post much. Teenagers spend most of their time in facebook which turns out to be a major means to state their opinion firmly, but adults are more involved in their commitment and show less interest in online activities.
Adult users are more likely to share their own photos, videosor share links of different incidents happening in society. Teenagers share the posts of entertainment pages, sports pages, celebrity photos, photo memes etc. There is notable difference in responses which teenagers and adult get for what they post on face book, as teenagers react more to their friends post. Number of friends is more for teenagers as they communicate and socialize with lot of people in school and college. These particular factors seem to be crucial in the research.

In this research, we focus only on two age groups teenager and adult, as they show significant difference in their behavior on social network platforms. Users between range 13 and 21 considered as teenagers, above 21 are considered as adults.

Algorithms for classification

There are many classification models which try to predict the values of one or more outcomes. Few machine learning algorithms for classification support vector machines, random forest, linear regression and artificial neural networks. However we do not confine to one model, but investigate on every model to get the best results. The Model which gives best accuracy is chosen by taking training time and iterations into consideration. Artificial neural networks show extremely good accuracy. And also training time is less in neural networks with the use of linearity. Other classification algorithms like support vector machines, multilayer perceptron exhibit moderate accuracy with average training time.

Classification is based on parameters extracted from user profile information. In the first stage parameters that influence the classification need to be identified. This can be done by analyzing user behavior and lots of data on facebook. In every problem of machine learning, accuracy depends on the parameters which are included in the analysis. Many users won't be active in facebook for days, these people have less information posted on their timeline, and these user samples need to be filtered out to reduce noise in the dataset.

In order to achieve desired results, analysis has to be done on large amount of dataset on machine learning algorithms like artificial neural networks and support vector machines etc., which may then help us to choose the best algorithm. Deep Learning Algorithms has gained lot of interest in recent due to its capability of predict with high accuracy. Deep learning has shown some outstanding results in classification problems.

In this research, we work on deep convolutional neural networks which outperform all other classification algorithms. Deep convolutional neural network is a deep feed-forward network with a variation of fully connectedneural network designed to reduce processing. DCNN has ability to approximate functions and perform classification tasks. DCNN is composed of different layers, where layers perform convolution operations and optimizations. Fig. 1 shows the architecture of the DCNN model with convolutional layer, subsampling layer and fully connected layer before the out-put. In the Fig.1 C represents convolution layer, S represents sub- sampling layer and M represents fully-connected multilayer perceptron.

The architecture of the model consists of two convolutional layers, two sub- sampling layers and fully connected layer at the end.

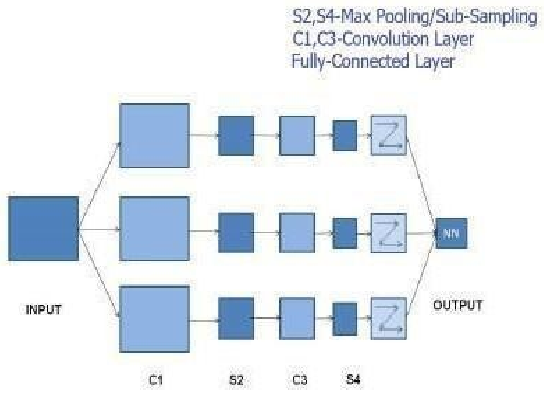

Fig 1. **Convolutional Neural Network Topology**

Age group classification model for twitter

The user profile information and user demographics are used to collect 7000 sentences which are qualitatively analyzed for classifying users into two broader sections of age group teenagers and adults. Hash tags, slang, punctuation, url, number of words in a post, number of users a particular user follows, number of followers for that particular user, number of tweets by that user, the topic of the tweet are the parameters considered to predict age group using twitter.

Machine learning algorithms such as Multilayer Perceptron, DCNN, Decision Tree, Random Forest, SVM are applied on the collected dataset for classification. [1]

3. AGE GROUPS CLASSIFICATION MODEL

WIn this section, a model is proposed for classifying age groups which includes two stages: 1) Data Extraction from facebook 2) Data Preprocessing 3) Classification

Data extraction from facebook

Before data extraction, number of posts which are directly posted by users on their timeline are analyzed to find out the parameters that are vital for classification task. There are many parameters, but among them only few parameters help in classifying age groups better. One among them number of friends of an user; teenager usually has more number of friends than an adult. Reaction for a post and style of writing of the user is also considered.

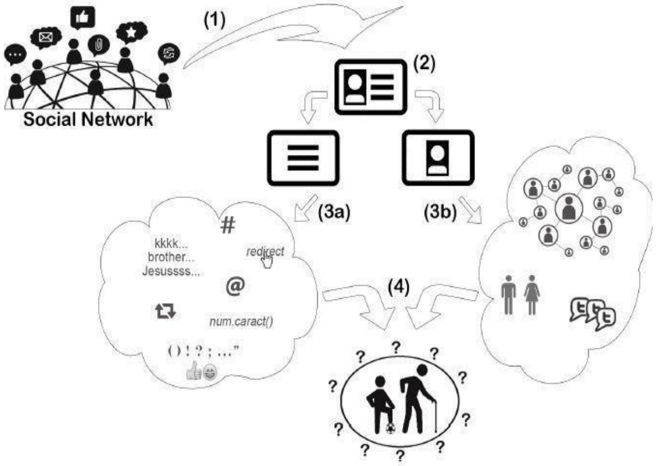

Fig 2. **Representation of proposed model**

A sample in the dataset consist of user profile information and his messages posted on his/her timeline. It is quite easy to analyze that teenagers post messages of longer length with more use of symbols called emoji's [4].

Teenagers are more expressive on social media platforms like facebook, so it is expected to have more number of posts on their timeline. In this research, we consider reaction for the messages posted on a timeline, where reactions include number of likes for the post, number of reactions for the post. Another parameter which can also differentiate an adult from teenager is use of hashtag '#'. It is often used at the end to tell about the topic or context of the message. Moreover, '#' is used to examine the topic trending on social media platforms. Parameters which are extracted from the user profile are number of page likes and number of friends. Pages share information about current happenings and issues in different areas. There are specific pages for entertainment, news, sports, cinema, food etc. User can 'like' any particular page if he/she wants to see the posts of that page in his newsfeed.

About 4000 samples of user profiles are collected and used in the classification task. All these samples were extracted using Facebook's Application Programming Interface (API). We go through the profile information and timeline of every user to extract the parameters from it. Whatever post appear on timeline is collected and total number of posts on user timeline along with reaction to the posts is counted.

A total number of 4500 samples are collected which contains user profile information and messages; about 520 messages were found as outliers after preprocessing. A total of 3980 valid samples were used for the data analysis. 80% of 3980 samples were used for training the model and 20% is used to validate the data.

Data preprocessing

All the samples collected in the first phase need to be preprocessed in order to ex- tract the useful parameters from it. Samples extracted only have messages and user profile information like number of friends, number of page likes in it. Parameters like length of the message posted, number of likes and number of comments are extracted from the post. Message text is processed to find the number of emoji's and use of hashtags in the message.And all parameters used in this research are numeric parameters.

Fig. 2 shows how samples are extracted and analyzed. Steps involved in first and second stage: 1) User profile information and messages posted on timeline is extracted from face book.
2)From each post, 2a) length of the message and other parameters from user profile were obtained 3) these parameters were analyzed using machine learning algorithms to predict the age group.

In this research, the following parameters were considered:
Message length: length of the message posted on timeline of a user

- Posts: total number of posts on timeline of a user
- Post likes: number of likes for a post or number of people reacted to the post.
- Comments: number of comments made by friends on a post
- Friends: total number friends or followers
- Page likes: number of pages a user has liked.
- Emoji's: symbols called emoji's used by user in a message to express his opinion.
- Hashtags: marker '# ' to tell the context of the message
- Label: this parameter represents the output of the algorithm that is teenager or adult.

Classification

All the relevant parameters to predict the age groups are defined in the previous section. Now these parameters are given input to the machine learning algorithms. Most of the parameters are numeric values there is no need to normalize the parameters, so it can be given input directly to the algorithm. The parameters as message length, posts, post likes, comments, friends, page likes, emoji's, hash tag are numeric and label is binary 0 or 1.

With all these parameters defined, machine learning algorithms were used and given input. Artificial neural networks show good generalization and the functions used for classification are approximated. The principle of artificial neural network is different layers extract features till it can predict with information propagated through the hidden layers simultaneously updating the weights to find the best hypothesis. Weights are learnt by propagating error backward through the network.

The test were implemented using anaconda framework and code for machine learning algorithms is developed in python programming language.

SVM or tree models are used in some studies but convolutional neural networks gave better results by extracting feature maps from the data using shared weights.

Following steps were involved in training of deep learning model for classification:
1) Design the architecture for the model 2) Input is given to the convolutional neural networks to train the model. We can also adjust the parameter of neural networks AS as shown in Table1 such as learning rate, batch size, decay factor and momentum. DCNN was trained for 50 epochs. Sigmoid classifier is used in the DCNN.

As stated, other machine learning algorithms like decision trees, multilayer perceptron and support vector machines are also used to predict the age groups. Decision trees group instances recursively by reducing the instability of classes. To do this, values of internal nodes are used and instances are positioned in leaf nodes. Therefore, non-leaf nodes compound to classification decision. Thus tree algorithm only acts as a classifier.

In this experiment, all the machine learning algorithms used same samples in order to evaluate how relevant the parameters are.

Table 1. Neural network parameters

PARAMETER	VALUE
Learning Rate	0.001
Momentum	0.9
Number of Hidden Layers	5
Number of Neurons in a layer	8
Epochs	100
Batch Size	20
Classifier	sigmoid

4. RESULTS AND DISCUSSIONS

In this section, the results of the proposed model for the classification of age groups are presented.

Classification in facebook

As mentioned in the previous sections, around 4000 samples are considered for the classification problem to predict the age groups. 80% of the total samples were used to train the model and 20% were used to evaluate the performance of the model.

In the training stage, DCNN obtained accuracy of 94.5% with Precision of 0.93 for teenagers and 0.92 for adults. In addition recall values reached 0.921 and 0.931 respectively. F-measure is also calculated which reached 0.930 for classification. Table 2 shows the results obtained by machine learning algorithms.

Parameters such as emoji's and hashtag do not influence much in the classification as its usage has become a common habit for both teenagers and adults.

Table 2. Machine learning results to classify age groups

PARAMETER	VALUE
Deep Convolutional NeuralNetwork (DCCN)	94.25%
Multilayer Perceptron (MLP)	92.00%
Random Forest	587.83%

Comparison of effectiveness of algorithms on dataset of Twitter and Facebook:

Classification of age group in Twitter using Deep learning is already analyzed and effective results were obtained using DCNN and MLP algorithms [1]. Table 3 shows precision, recall and F-measure for twitter dataset based on the experiments conducted in reference project using twitter dataset[1].

Table 3. Precision, Recall and F- measure of Twitter Dataset

Algorithm	Precision	Recall	F-measure
DCNN	0.92	0.93	0.93
MLP	0.88	0.86	0.88

In recent years many social networking platforms were developed, but facebook stands to invincible with no other competitor nowhere near to it. Therefore age classification in face book is essential as it is proved that facebook has tremendous impact on its audience unlike other social networking platforms. Hence this paper focuses on classification of age group in Face book using Deep Learning. Table 4. shows precision, recall and F-measure for Facebook dataset based on experiments and algorithms stated in this paper.

Table 4. Precision, Recall and F- measure of Facebook Dataset

Algorithm	Precision	Recall	F-measure
DCNN	0.94	0.94	0.94
MLP	0.91	0.71	0.80

5. CONCLUSION AND FUTURE WORK

AUser demographics like gender, age, educational level and personality can be learned by mining the massive volumes of personalized and diverse data produced in public social media. In this research we focus on the prediction of the age information that is to classify people into two labels such as, teenagers or adults because this information sometimes is not available but that information is needed for recommendation systems.

Attributes such as number of emoji's and number of hashtags are very specific to age group as it can be inferred that teenagers use more number of emoji's or hashtags when compared with adults thus these attributes ensure higher degree of distinction. According to our results it is also observed that the message length, number of posts, number of likes and comments on them also guarantees reliability to results. Also the parameters emoji's and hashtags did not influence the performance of the classification model.

Considering several facebook user's profile and user's history about 4500 sentences were analyzed to obtain the parameters that are required for the research. The accuracy obtained makes it evident that the parameters used in the research are highly reliable.

The DCNN machine learning algorithm had an accuracy of about 94.25% outperforming Multi Perceptron (92.00%) and Random Forest (87.83%) and thus is chosen to be the best way for classifying age in social networks.
Facebook where some of the personal information of the user like age might not be available the research provides a quite accurate method through which the user age can be classified under two labels thus providing more personalized experience for the user and enhancing the recommendations for that particular user depending upon the label they fall into.

Our future work includes extending the age classification model to detect and prevent the various types of digital misuse being done with the social media. Furthermore, the co-author of this paper has also developed a data fusion based digital investigation model [10] [11] which can address various types of cyber threats in network and facilitate computer forensic analysis. In addition to the various applications of data fusion methodology in military and nonmilitary areas and digital forensic analysis; it can be effectively extended to related areas such as sentiment analysis and age classification in social medias for data integration, preprocessing, pattern classification. Furthermore the model being proposed here can be integrated with various existing computer forensic tools to enhance the accuracy and efficiency of digital forensic investigation process and help the law enforcement agencies to detect and prevent various types of computer frauds and cyber crime; hence maintain the digital technology as a platform for wellbeing of the nation.

6. REFERENCES

[1] Rita G. Guimaraes, Renata L. Rosa, Denise De Gaetano, Demostenes Z. Rodıguez, Grac¸a Bressan, "Age Group Classification in Social Network Using Deep Learning", vol. 5 Publication Year: 2017, Page(s):10805 – 10816.

[2] Rebecca Anhang Price, PhD, Marc N. Elliott, PhD, Alan M. Zaslavsky, PhD, Ron D. Hays, PhD, William G. Lehrman, PhD, Lise Rybowski, MBA, Susan Edgman-Levitan, PA, and Paul D. Cleary, PhD., "Examining the Role of Patient Experience Surveys in Measuring Health Care Quality".

[3] Crime pattern detection using online social media Raja Ashok Bolla.

[4] Aindrila Biswas, Nikhil Ingle1 and Mousumi Roy, "Influence of Social Media on Voting Behavior"

[5] Zheng Chen, Xiaoqing Du, "Study of Stock Prediction Based on Social Network".

[6] Anton Alekseev,Sergey Nikolenko, "Word Embeddings for User Profiling in Online Social Networks".

[7] Wiesław Wolny, "Sentiment Analysis of twitter data using emoticons and imoji ideograms".

[8] Q You, S Bhatia, T Sun, J Luo, "The eyes of the beholder: Gender prediction using images posted in online social networks".

[9] Jalal S Alowibdi UgoA. Buy Philip Yu, "Language independent gender classification on Twitter."

[10] Suneeta Satpathy, Sateesh K. Pradhan, B.N.B Ray, “A Digital Investigation Tool based on Data Fusion in Management of Cyber Security Systems”, International Journal of IT & Knowledge Management, 3(2), pp561-565,June2010.http://www.csjournals.com/IJITKM/PDF%20 3-1/77.pdf.

[11] Suneeta Satpathy, Asish Mohapatra, “A Data Fusion based Digital Investigation Model as an Effective Forensic Tool in the Risk Assessment and Management of Cyber Security Systems, 7th International Conference on Computing, Communications and Control Technologies (CCCT), July10- 13,2009,Orlando,USA.