Developing Online Age Verification Tools Using Supervised Machine Learning Models

Clara Arnold
Boise State University
Boise, USA
claraarnold@u.boisestate.edu

Aidan Flinn
Boise State University
Boise, USA
aidanflinn@u.boisestate.edu

Md. Mashrur Arifin
Boise State University
Boise, USA
mdmashrurarifin@u.boisestate.edu

Dr. Jyh-haw Yeh
Boise State University
Boise, USA
jhyeh@boisestate.edu

ABSTRACT

We want to develop an online age verification tool to identify the age group of a user in a non-invasive way using supervised machine learning tools. We present a manually created data set that has over 500 user profiles with different features identified to help in this task. We have used Google Collaborator to train and test three supervised machine learning models: support vector machines, decision trees, and random forests. The data set has proved to test with over 85% accuracy for all three machine learning models. We further separated the data set to test the "certain" and "uncertain" features to get more meaningful results and found that the "certain" features increase the accuracy.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; *Social networking analysis*; • **Computing methodologies** → Supervised Machine Learning.

KEYWORDS

Age verification, social media features, supervised machine learning tools, public information

ACM Reference Format:
Clara Arnold, Aidan Flinn, Md. Mashrur Arifin, and Dr. Jyh-haw Yeh. 2023. Developing Online Age Verification Tools Using Supervised Machine Learning Models. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Most online age verification tools require users to simply hit "yes" or "no" indicating they are above the specific age that an age-gated site might need, allowing users an easy out if they are not the required age. Some age-gated sites require users to prove their age with government-issued records which is an invasive method to validate a user's age. We want to create a tool using supervised machine learning that will allow all users to confirm their age in a simple, non-invasive way.

To do this we have manually created a data set of 569 public Instagram profiles and determined different features that will help the three machine learning tools used determine the age of users based on publicly available information. The features include an explicitly typed age, the school a user currently attends (whether that be a middle or junior high school, a high school, or university), sports the user participates in, graduation captions, captions that contain information about the current grade the user is in, captions about school dances, the number of hashtags included in the profile, the number of photos the user is tagged in, the amount of followers, and the amount the users the profile is following. To manually create this data set we first went through many public profiles and determined their age using the features defined above, the age groups identified are as follows: Under 12, 13-17 and Over 18.

We chose to use supervised machine learning models because our data set contains labels where the input data corresponds to the output data or target. We wanted the models to be able to use the output and input included in the training data set to accurately be able to map unseen output to a target age group. Further, we chose support vector machines, decision trees, and random forests to train and test our data set. We did not choose linear regression as many other researchers have because not all of our data is binary, many of the features require unique values such as "Prom" in the dances feature or "Basketball" in the sports features. We also did not choose neural networks because we didn't want to use deep learning for our data set just yet, we wanted to see how it was performing before diving into this type of training. However, we did choose support vector machines because this model is known for being able to separate classes well and since we have three target classes we thought this would help with the accuracy scores. We also did choose decision trees and random forests because they can interpret data well and provide higher accuracy and robustness.

Further, we tried many different combinations of the data set to see which features were the most important in training the machine learning models to identify a users age. The main focus of this was the see if the "certain" features, which are the explicit

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

age and mention of school a user currently attends, affects the accuracy the machine learning models provide. We also tried many combinations of the "uncertain" features, the remaining eight, to see which of these were found to be weighted the most heavily during the training and testing process.

This research is important because it will prevent minors from having inappropriate internet access and allow vendors to ensure their users are of age. To emphasize, our main goal in this research is to create a non-invasive, interactive, and accurate tool to identify a user's age online.

2 RELATED WORK

Through our research we came across many different approaches to solving this problem at hand, verifying age in a non-invasive way. A very popular approach found was using facial recognition tools to determine the age of an individual based on a photo. Andreas Lanitis et al. [3] tested how well different classifiers can accurately estimate the age of someone based on a facial image by analyzing 22 different facial features and found that hierarchical age classifiers improve the accuracy of age estimation for all methods and the best results occurred when age-specific and appearance-specific classifiers were combined. In similar research, Hu Han et al. [4] used a binary decision tree based on a support vector machine to automatically estimate the age of an individual by using a component-based representation technique by using a method based on ESM to localize individual facial components to detect a set of landmarks followed by component localization based on subsets of landmarks. They then extracted the BIF features from the individual facial components and found the FG-NET data set performed best when all the three best features of holistic BIF, shape, and eye region BIF were combined, and they also found that the MORPH Album2 and PCSO data sets had the best results with the holistic BIF features. Another fascinating paper we found which also used facial recognition tools to determine age looked more closely into the facial aging pattern that human faces follow. Xin Geng et al. [5] created a tool named AGES that uses aging patterns as a sample rather than individual images by using PCA to construct subspaces of variation in facial images and using LDA to deal with that variation. They found that AGES performed better than the most commonly used algorithms for this type of problem. From all the papers we read that used this approach to identify age in a non-invasive way we decided there wasn't much we would be able to contribute to this specific research.

We continued to read a number of research papers in order to come to a conclusion about where research was needed for this topic and how we could make a valuable contribution to the research community, specifically in age verification and machine learning. From all that, we learned quite a lot regarding machine learning, data collection and how others have tried to verify age in non-invasive ways which lead us to our conclusion of collecting public information and training machine learning models to see how accurate we could get them.

After coming to this conclusion we began looking into common keywords that different age groups commonly use and found that H. Andrew Schwartz et al. [6] had done extensive research on keywords relating to age, gender, and personality. They used an open-vocabulary technique which allowed them to find connections that aren't traditionally captured in this type of keyword research and found a progression of school, college, work, and family topics when looking at the dominant topics of each age group which allowed them to conclude that typical concerns peak at different ages. We had initially intended to implement this keyword based verification through using an internet search including a person's name and email to compile a list of URLs to then read through the text information on those websites to then reference a list of keywords that would determine the age of the person searched. This method had increasing variability due to the immense amount of information on the internet and possible duplicate names/surnames that could cause confusion. As well as many people not having much information available on the internet such as young children and elderly people. We had decided to look into social media for the information to be gathered. Antonio A. Morgan-Lopez et al. [7] looked at the language of users on Twitter and found that vocabulary, writing styles, and speech patterns endure over time as individuals learn and develop. They collected data regarding birthday tweets and created models using four variable sets: language features only, metadata features only, language and metadata features, and words/phrases. These keywords and ideas of commonalities between topics and age helped us in determining what features we wanted to include in our data set, which is talked more about in the next section.

2.1 Collecting Features

Before we did any outside research looking for features we simply tracked the different aspects of the profiles we were looking at to see the common posts or common words used that might help us determine age. We found that many people post their actual age such as when it is their birthday and almost all users have some implication of where they currently go to school or used to go to school. Many users also had some type of sport which gives idea to their age, a graduation post and even captions or tags about what grade they are currently in.

From there we looked into published research and found that Kyungsik Han et al. [1] created two independently created data sets which focused on profile-based age detection and tagged-based age detection. The machine learning models they used were able to test with 82% accuracy which is what we aimed to beat in our research since the ideas of our research are so similar. From this research though, we were able to gain a few new feature ideas including the number of hashtags within a profile as well as tracking the number of followers/following.

From another article, Junho Song et al. [2], we learned that many users caption their selfies with certain tags that can identify the post as a selfie without any facial recognition tools. This project reported a 88% accuracy using linear regression machine learning tools and a 74% accuracy using random forests.

3 DATA COLLECTION

We manually created a data set containing 569 user profiles over the course of nine weeks. Originally, we had planned on collecting data on people through a public internet search but realistically collecting a data set for this would not have been possible so we

turned to Instagram. Each profile was randomly chosen to be part of the data set and only publicly available information is recorded in our data set. There are 42 profiles in the "Under 12" target, 279 profiles in the "13-17" target, and 248 profiles in the "Over 18" target.

The data set has a two "certain" features and eleven "uncertain" features with a total of thirteen features. To specify, a "certain" feature is a feature in which the exact age or target age group is given and highly accurate. For example, if a user posts about their birthday with a "Finally legal #18" caption the user is 100% in the "Over 18" target group. An "uncertain" feature is a feature in which the user's age is not given and we are simply collecting data on the user in order to be able to gain some type of trend from the data collected. For example, those in the "Over 18" target tend to have more followers and follows than those in the "Under 12" target. As shown below and in the Results section, we trained and tested the supervised machine learning models with several different combinations of the "certain" and "uncertain" features.

As for the training and testing of the three machine learning models we chose to perform this project. Our manually created data set of 569 samples we split 70-30 where 70% of the data set was used for training and 30% was used for testing. To reiterate, the three machine learning models we chose are the support vector machine, decision trees, and random forests. We chose these models because they allowed us to analyze how our data set was doing as well as try out different combinations of features to reach the highest accuracy's possible with the data set created.

3.1 Data Analysis

Over the course of the development of our data set, we tested the effect of many features on the several machine learning models. Initially our data set only had a few features such as if a profile contained an exact age or place of schooling. These two features produce a higher success due to the immediate correlation between the feature and the proposed target age. We began increasing the amount of features capable of being taken from a users profile that do not exactly or greatly suggest the age of the user. For example some of these these non age related features include the number of posts, followers/following, and hashtags/tagged posts. By including a greater amount of features not relating to age we could gather more significant data from our models.

Another large development in the effectiveness of our data was reducing the variability of the data points in each feature. Initially when plotting most of the data we gave many points unique values taken directly from the user profile. This means each different unique value was encoded to a different value in the machine learning model. Which even if the values were similar values (for example class of 2025 vs C/O 2025) they would be seen as separate values and not helpful in mapping the data by the model. Over time we went through the data set to make all of the features format the values in a similar way so that the machine learning model tracks them as the same value.

Throughout our progress on this research we added more and more features to increase the accuracy of the models but also to spread out the significance of each feature. With more features, the more factors affect the data and the more accurate it becomes. Initially with this project we had 5 features. Age_exact, School_name, Sports_tag, Grad_tag, and Grade_tag. Through development we added many more features which would also allow for more data to be plotted and increase the overall accuracy of each model.

4 RESULTS

As stated above in the Data Collection section, we tried many different combinations of features to see which combination produced the highest accuracy as well as trying to features individually to see which feature provided the most information for the machine learning models. When we included the "certain" features, the exact age and the current school of the user, the accuracy for the machine learning models can be seen in Table 1.

Table 1: Accuracy in Absence of "Certain" Features

Model	Age_exact	School Related Features	All Features
SVM	69	85	87
DT	74.5	86.8	94.7
RF	80.7	91.2	96.5

In this table it can be seen that the use of the school related features (Middle_or_junior_high, High_school, University, or School_name) increased the accuracy of the support vector machine and decision tree machine learning models, but the random forest's accuracy remained stable. Overall, the use of both the exact age feature and current school feature increased the accuracy the most which is the highest accuracy for all three models we saw with this data set and these models.

When we dropped the "certain" features and tried out the "uncertain" features which include features that don't directly relate to age we saw a decrease in accuracy. The support vector machine performed at 69% accuracy, the decision tree performed at 74.5% accuracy, and the random forest performed at 80.7% accuracy. These results are not better then results in previous research so we continued on to test and see which of the features was producing the best information for the machine learning models.

4.1 Individual Feature Performance

We went through and dropped each feature one at a time to see which feature had the biggest weight on the accuracy produced. Table 2 shows the results where the feature listed was NOT included in the training or testing data set.

To specify all the features in the table, profile name is the username of the user, Age_exact is if the user provided their exact age within their profile, School_name is if the user provided the current school they are attending such as "JMS" or "BHS" or "UO", Sports_tag is if the user plays a sport, Grad_tag is if the user provided a graduation post whether is be high school or university, Grade_tag is if the user specified the grade they are in such at "7th grade" or "Senior year", Dance_name is if the user posted about a high school or university dance, Hashtag_amount is the amount of hashtags the user has on their profile, Number_of_reels is the number of videos posted on the profile, Number_of_photos is the

Table 2: Accuracy Without Individual Feature

Removed Feature	SVM	DT	RF
Profile Name	87	96.5	96.5
Age_exact	85	86.8	91.2
School_name	88	93.86	96.5
Sports_tag	85	95.6	96.5
Grad_tag	87	93	96.5
Grade_tag	88	90.3	94.7
Dance_name	87	97.4	96.5
Hashtag_amount	84	94.7	95.6
Tagged_photos	87	94.7	96.5
Number_of_reels	86	93.8	95.6
Number_of_photos	85	94.7	96.5
Followers	85	94.7	96.5
Following	87	94.7	94.7

number of photos posted on the profile, and followers/following is the amount of followers and follows the user has.

Based on Table 2 it is clear to see that the Age_exact, Sports_tag, Hashtag_amount, Number_of_reels, Number_of_photos, and Followers caused a decrease in accuracy for the support vector machine when taken out of the training and testing model. The largest decrease happened when the Hashtag_amount was taken out of the data set which implies that this feature has a great impact on the support vector machine. For the decision tree, the Age_exact, School_name, Grad_tag, Grade_tag, and Number_of_reels caused a decrease in accuracy with Grade_tag caused a 7.4% drop from the training and testing data set with all features. This implies that the Grade_tag feature has importance in training and testing for the decision tree model. Finally for the random forest model, the Age_exact, Grade_tag, Hashtag_amount, Number_of_reels, and Following tags caused the greatest decrease in accuracy. The Age_exact tag being removed caused a 5.3% decrease in accuracy from the original data set meaning the exact age of a user being provided increase the accuracy of the random forest model.

5 DISCUSSION

We aimed to see if we could create an effective tool through using Instagram for data collection and supervised machine learning models for training and testing the accuracy the data set provides. Through this research we definitely found trends within the three different age groups we aimed to identify, Under 12, 13-17 and Over 18. While there were many limitations and much future work can be done to create tools to identify these particular age groups, our research contributes to this problem because we have identified and tested new features with high accuracy's that previous research has not yet accomplished.

5.1 Limitations and Future Work

Since the data set is a manual compilation there is a limitation in the amount of data points that can be compiled realistically in a 9 week period. We also were limited to the amount of features we can utilize without another model to distinguish several more features from Instagram profiles. There is also inherent bias with utilizing social media as usually some profiles have more information than others. The data set could not be compiled without some way of verifying the age of a user, there are many profiles on social media sites that have relatively low information or none at all pertaining to the user so they are unreliable to use. Utilizing users with little to no features or no way of verifying age would give no meaningful data to the machine learning model and could possibly skew trends in accuracy. There are limitation to using Instagram as when a profile is private if one is not actively following said user, then the viewed private profile's photos and posts are not visible. Applying the goal of this research to another social media platform with more abundant information available to all users regardless of private status could provide more accurate results.

6 CONCLUSION

When we trained and tested our manual data set which included all features both "certain" and "uncertain" using the Random Forest supervised machine learning model we saw the highest accuracy of 96.5%. This research contributes to determining the age of someone based only on their Instagram and we hope to continue or see future work that uses the features identified in this paper to look at trends on other social media accounts and even just the public internet.

7 ACKNOWLEDGMENTS

This research was supported by National Science Foundation Cloud Computing and Privacy REU site under award number 2244596 at Boise State University and the Department of Computer Science of Boise State University. We would also like to thank our mentor Dr. Jyh-haw Yeh, as well as Dr. Jerry Fails for their support and guidance with this research.

8 REFERENCES

[1] Han, Kyungsik. Lee, Sanghack. Jang, Jin Yea. Jung, Yong. Lee, Dongwon. (2015) *"Teens are from Mars, Adults are from Venus": Analyzing and Predicting Age Groups with Behavioral Characteristics in Instagram.* Pacific Northwest National Laboratory, USA. The Pennsylvania State University, USA.

[2] Song, Junho. Han, Kyungsik. Lee, Dongwon. Kim, Sang-Wook. (2018) *"Is a picture really worth a thousand words?": A case study on classifying user attributes on Instagram.* Department of Computer Science and Engineering, Hanyang University, Seoul, Republic of Korea. Department of Software and Computer Engineering, Ajou University, Suwon, Republic of Korea. College of Information Sciences and Technology, Pennsylvania State University, University Park, United States of America.

[3] Lanitis, Andreas. Draganova, Chrisina. Christodoulou, Chris. (2004) *Comparing Different Classifiers for Automatic Age Estimation.* Department of Computer Science and Engineering, Cypress Semiconductors, Inc., Nicosia, Cyprus. Department of Computing, Communication Technology and Mathematics, London Metropolitan Univerty, London, UK. School of Computer Science and Information Systems Birkbeck College, University of London, London, UK.

[4] Han, Hu. Otto, Charles. Jain, Anil K. (2013) *Age Estimation from Face Images: Human vs. Machine Performance.* Department of Computer Science and Engineering Michigan State University, East Lansing, MI, U.S.A.
[5] Geng, Xin. Zhou, Zhi-Hua. Smith-Miles, Kate. (2007) *Automatic Age Estimation Based on Facial Aging Patterns.* School of Engineering and Information Technology, Deakin University, VIC, Australia. National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. School of Engineering and Information Technology, Deakin University, VIC, Australia.
[6] Schwartz, H. Andrew. Eichstaedt, Johannes C. Kern, Margaret L. Dziurzynski, Lukasz. Ramones, Stephanie M. Agrawal, Megha. Shah, Achal. Kosinsk, Michal. Stillwell, David. Seligman, Martin E. P. Ungar, Lyle H. (2013) *Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.* 1 Positive Psychology Center, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America.Computer & Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America. The Psychometrics Centre, University of Cambridge, Cambridge, United Kingdom.
[7] Morgan-Lopez, Antonio A. Kim, Annice E. Chew, Robert F. Ruddle, Paul. (2017) *Predicting age groups of Twitter users based on language and metadata features.* Behavioral Health and Criminal Justice Research Division, RTI International, Research Triangle Park, North Carolina, United States of America. Center for Health Policy Science & Tobacco Research, RTI International, Berkeley, California, United States of America. Center for Data Science, RTI International, Research Triangle Park, North Carolina, United States of America.

Received 11 August 2023