Using Machine Learning & AI to tackle Online Hate and 'Anti-X' Sentiment

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Richard Sear (Sophomore undergrad, Comp. Sci. & Physics), Nicolas Velasquez (post-doc, Elliott School), Neil Johnson (Faculty, Physics)

Research Question

Can an algorithm automatically classify social media users that are members of online hate groups, based solely on the public image that they chose to represent themselves in their online profile?

Our motivation: Social media companies, agencies and governments are struggling with the task of identifying and removing online extremist content, including hate speech -- and in particular, whether specific groups of people are promoting it.

Any success in this area will help open the way for a scheme in which AI continually interacts with image-based social media, to detect the development of contentious societal issues that have potentially harmful real-world consequences.

In addition to extreme/hate speech surrounding gender, religion, and race, other potential applications of AI range from anti-vaxx and other areas of public health, through to national security risk assessment. Also of interest for this research are harmful images aimed at denigrating consumer products and public figures (possibly including deep fakes).

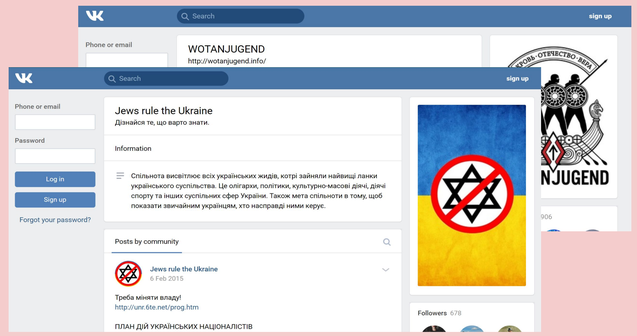
Project Implementation

We use social media data from online hate groups on the VKontakte social media platform, which is known for its extremist content [1,2] (www.vk.com). Our dataset contains 5,331 profile images, all of which are publicly available. We have no need of, nor do we collect, any private information about individuals. An additional 4,555 faces came from the AffectNet dataset and were used as training data for faces not in an online hate group. We trained a multilayer classifier using the Microsoft Cognitive Toolkit module for Python. Of the total 9,886 images, 6,920 were used as training data, 1,483 were used as test data, and 1,483 were used to evaluate the accuracy of the trained model.

Group 0

Group 1

WARNING: These are screenshots of actual hate material found online



All our data is 100% publicly available information, comprising the public pictures that individuals choose to represent themselves online. We scraped our Group 1 images from pages containing overtly hateful content. Our question is whether their public profile pictures offer predictive power for whether they end up in an online hate group.

Outline of our CNN

Image pixel data (50 x 50)

7 hidden layers

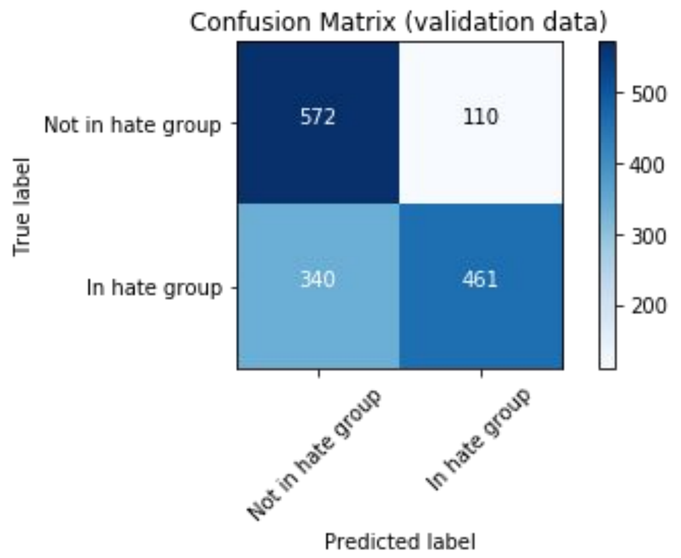
Classifications (0 or 1)

Our machine learning classifier attempts to correctly identify as many objects pictured in an image as possible. We use a Convolutional Neural Network (CNN).

Results

Our results show a surprisingly high success rate for the CNN (approximately 70%), in terms of predicting whether a given facial profile belongs to a member of an online hate group. A random classification yields only a 50% success (selecting between two options). We are now exploring the likely reasons for this success, as well as providing some initial observations concerning broader 'anti-X' hate aimed at historically marginalized populations, religion, race, women and genders -- as well as identifying deep fakes.

We conclude that a CNN classification algorithm can indeed be used to classify social media users that become members of online hate groups, based solely on the avatar that they choose to represent themselves in their online profile.



We are grateful to Rhys Leahy and Nicholas Johnson Restrepo for assistance with the data collection.

[1] "Social media cluster dynamics create resilient global hate highways"; N.F. Johnson, R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique. Nature (under review). Available at https://arxiv.org/abs/1811.03590
[2] "New online ecology of adversarial aggregates" Science 352, 1459 (2016); N. F. Johnson, M.Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, S. Wuchty