

# A Big Data Science Experiment Protecting Minors on Social Media Platforms

Estée van der Walt<sup>1</sup>, J.H.P. Eloff<sup>2</sup>  
estee.vanderwalt@gmail.com<sup>1</sup>, eloff@cs.up.ac.za<sup>2</sup>



Cyber-security and Big Data Science research group, Department of Computer Science, University of Pretoria, South Africa

## Overview

- Many people present a false identity for various purposes, whether for anonymity or for something more harmful like pedophilia or terrorist activities.
- The big data characteristics of social media make it not only easier for people to deceive others about their identity, but also harder to prevent or detect identity deception.
- The research propose a so-called Identity Deception Score (IDS to assist law enforcement with the early detection of potentially harmful behavior on social media platforms.

## Methods

### Gathering data

- Gather social media data from Twitter.
- Gather tweets from a similar population.
- Include the network of the population (friends and followers).

### Data cleanup and imputation

#### Remove

- Retweets
- Bots and celebrity accounts
- Closed or new accounts

#### Add

- Known deceptive accounts

### Attribute selection

- Use SOMs, network diagrams, clustering and other EDA techniques to evaluate all attributes available.
- Identify those attributes leading to deception.

### Feature engineering

- Friend/follower ratio
- Avg. tweet time
- Number of devices
- Distance between given and perceived location
- Sentiment
- Image type

### Supervised machine learning

- Build various models to detect identity deception
- Evaluate the results from these models

### Building an identity deception score

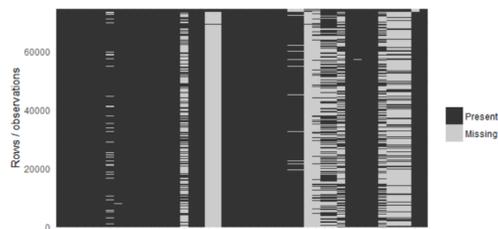
- The IDS is an intuitive representation of identity deception using the results from supervised machine learning models

## Results

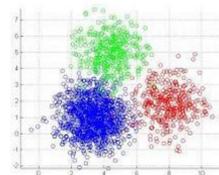
606,914,240 tweets  
+- 4K per tweet



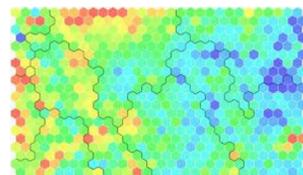
223,796 unique users



Network diagram



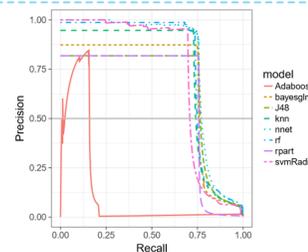
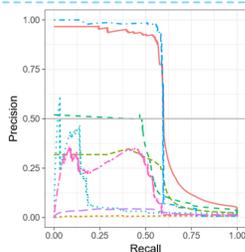
Clustering



SOM



Bagging of words



IDS = ???

## Architecture



Twitter



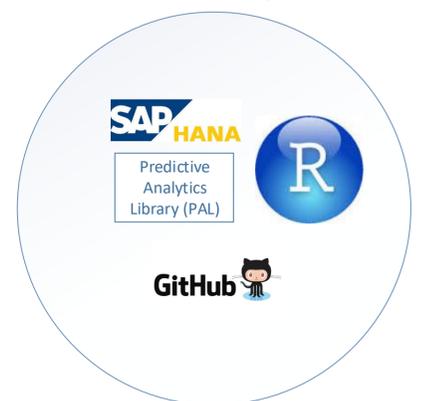
hadoop



Latitude + Longitudes



SAP HANA



## Conclusion

- The research at hand identified and evaluated various features that could play a role in identity deception on a social media platform.
- It was found that engineered features previously used to detect non-human accounts did not perform well.
- It was found that engineered features built from knowledge in social sciences (psychology) could predict deception quite well.
- The results are difficult to explain due to the nature of machine learning models (usually black box models).

### Future work:

- Determine which attributes or features contributed the most to identity deception during the previous supervised machine learning experiments.
- Build a simple, intuitive algorithm to score each user account knowing the contribution mentioned before. This score will be known as the Identity Deception Score (IDS).
- Explain how the IDS compares to machine learning and what the advantages is from one approach vs another.