A Big Data Science Experiment Protecting Minors on Social Media Platforms

Estée van der Walt[1], J.H.P. Eloff [2]
estee.vanderwalt@gmail.com[1], eloff@cs.up.ac.za[2]

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 to remain anonymous or for some other malicious purpose, like online grooming.
- 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 proposes to use features from social sciences to detect identity deception.
- Lastly the results are presented in the form of an Identity Deception Score (IDS) of which the results are transparent and explainable.

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

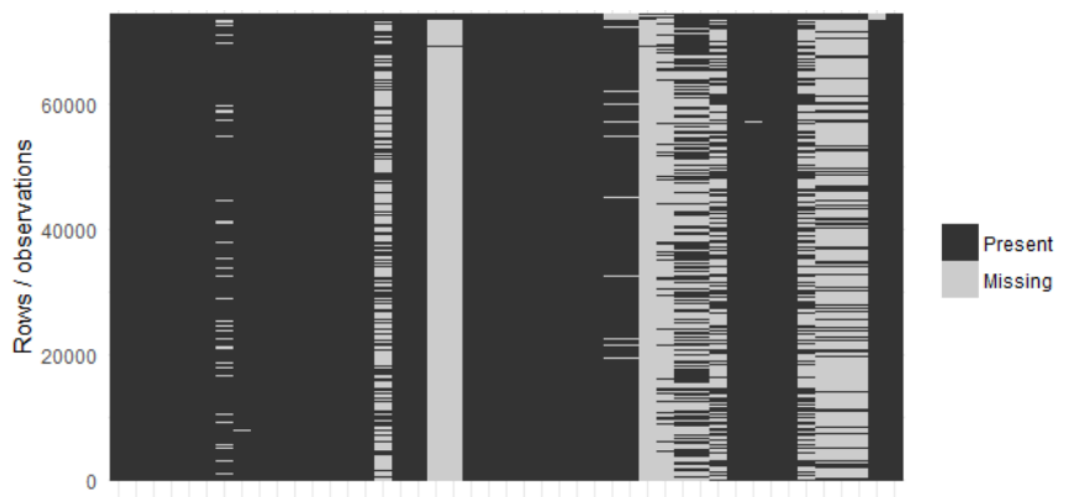
Building an identity deception score

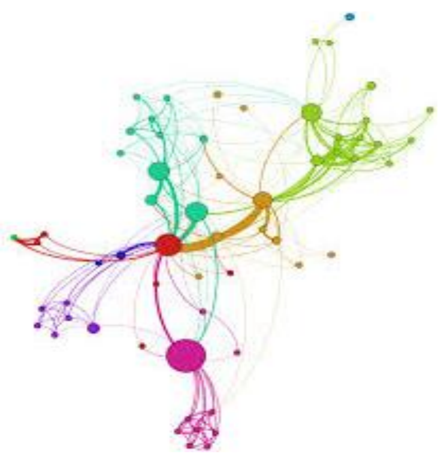
- The IDS is an intuitive representation of identity deception using the results from supervised machine learning models

Results

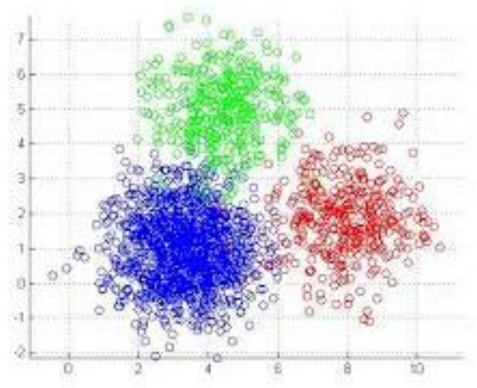
606,914,240 tweets
+- 4K per tweet

223,796 unique users

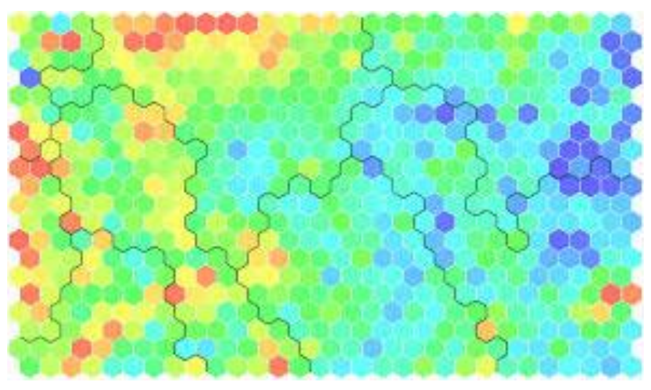




Network diagram

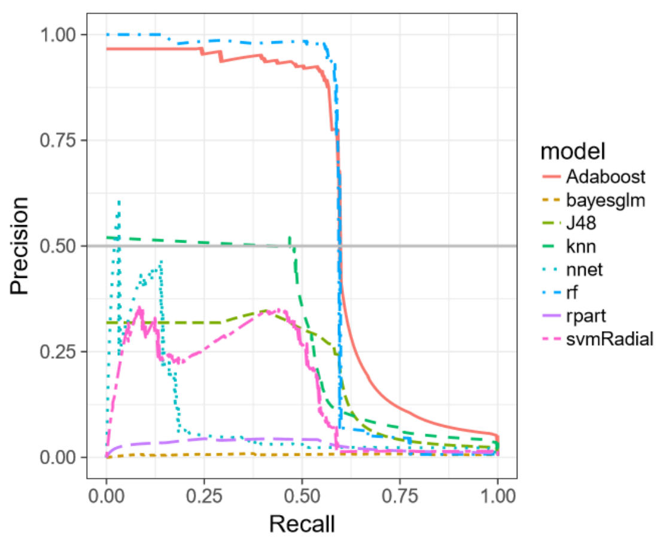


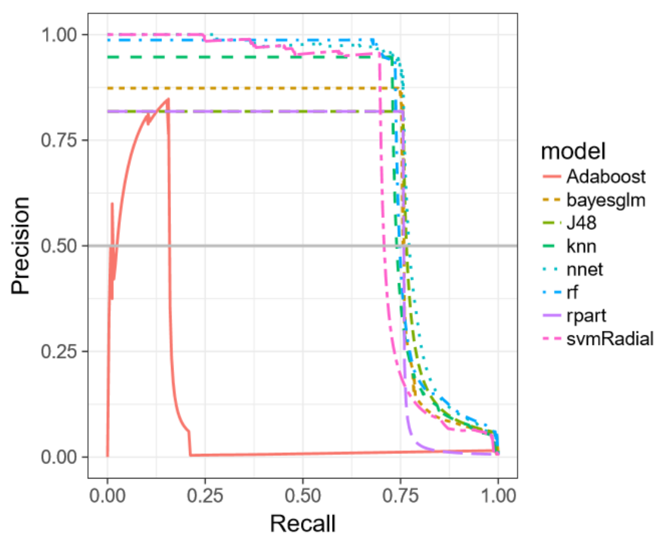
Clustering



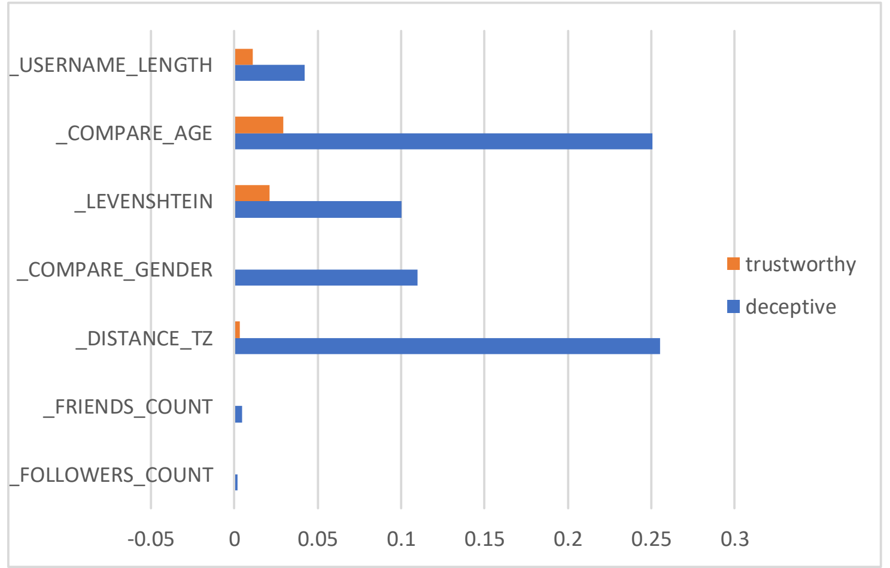
SOM

Bagging of words

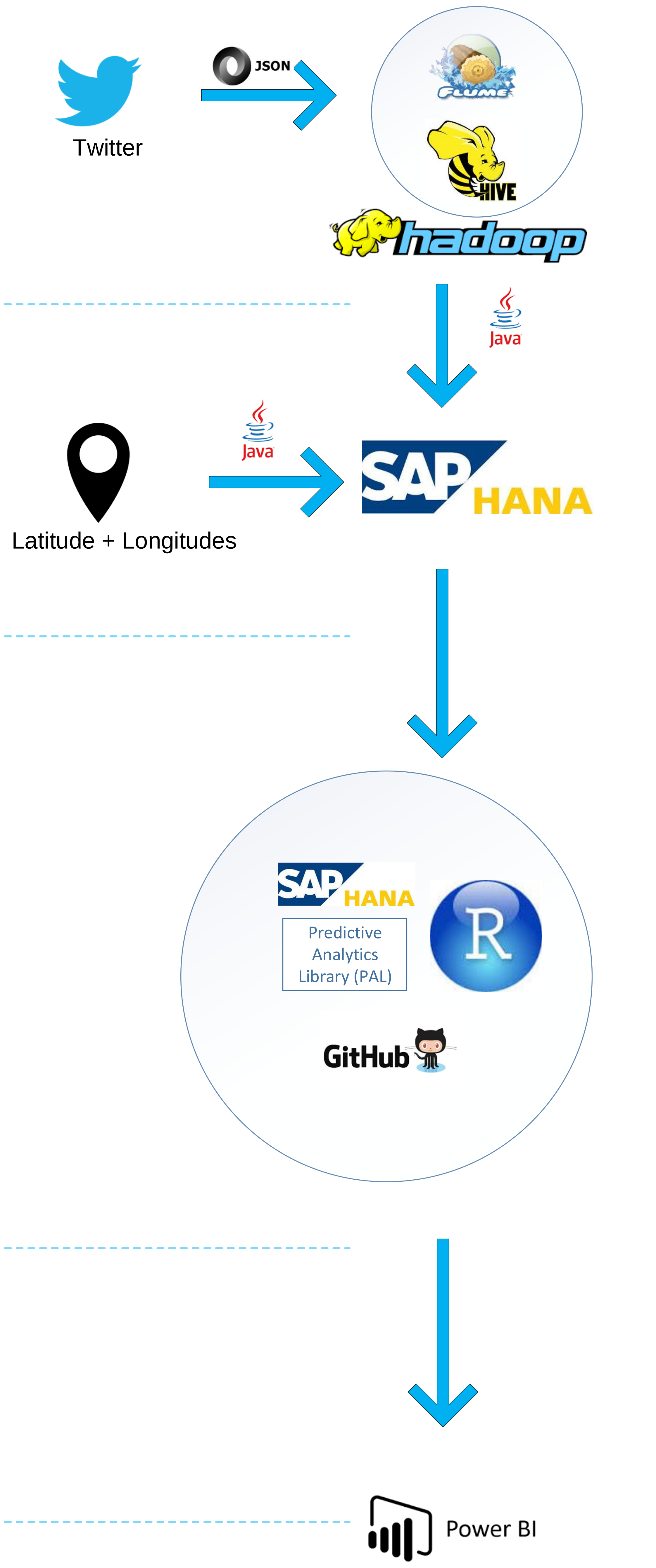




IDS = ???



Architecture

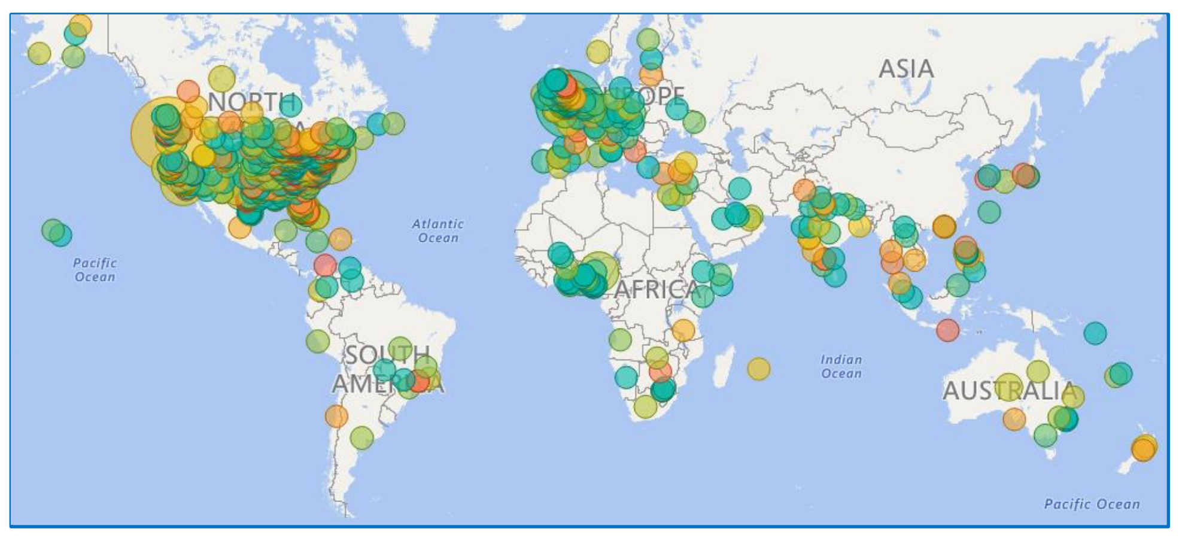


Twitter

Latitude + Longitudes

Predictive Analytics Library (PAL)

Power BI



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 (bots) did not perform well to apply directly to humans.
- It was found that engineered features built from knowledge in social sciences (psychology) could better the prediction of identity deception considerably.
- 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).
- Use the IDS to explain the results in a more intuitive way than current black box machine learning models.