ISSN (Print): 0974-6846
ISSN (Online): 0974-5645
Indian Journal of Science and Technology, *Vol 12(29), DOI: 10.17485/ijst/2019/v12i29/146977, August 2019*

Cambridge Analytica – A Case Study

Harshil Kanakia*, Giridhar Shenoy and Jimit Shah

Sardar Patel Institute of Technology (Autonomous Institute Affiliated to University of Mumbai), Mumbai - 400058, Maharashtra, India;
email2harshil@gmail.com, giridharpshenoy@gmail.com, shahjimitindia@gmail.com

Abstract

Objectives: This study discusses how Facebook users' data has been harvested, used to formulate an algorithm to understand users' personality traits and in-turn use the process to influence the outcome of US Presidential Elections. **Method:** A Quiz application was developed to collect Users' data. Their activities on Social Media were analyzed, patterns were detected, OCEAN scores were given, and user groups were made on the basis of their political affiliation. And finally, they were targeted with suitable ads and news to achieve desirable results. **Findings/Application:** Use of Internet is increasing day-by-day and so is our Digital Footprints. Companies like Netflix started using these data to understand their Customers' behaviour and to improve their Customer Oriented Marketing Strategies. It has been observed that activities on Social Media say a lot about the users and sometimes it is used by companies harvesting user data in unethical ways.

Keywords: Cambridge Analytica, Facebook, OCEAN score, US Elections

1. Introduction

The use of information to manipulate people is an old concept referred to as 'theory of mind[1-2]. Though this concept has various perceived uses, an important use in today's socially connected world is to deceive. Deception is an evolutionary trait in some animals E.g. Camouflage or acquired gradually by learning in case of humans. This combined with the human intelligence makes the level of deception sophisticated or rather scary.

This study deals with a recent scandal involving Cambridge Analytica, Facebook and the US Elections, where known information was used to deceive and manipulate people in order to change their political views and votes.

Cambridge Analytica was found to be using Facebook Data sourced from a Cambridge University professor to work for a US Presidential Candidate. This malpractice was exposed by the former director of Cambridge Analytica, Christopher 'Chris' Wylie. The estimate of user data breached ranges anywhere from 30 million ~ 80 million profiles.

2. Literature Survey

According to[3] Cambridge University professor built a Facebook Quiz app that exploited a loophole in Facebook API that allowed collecting the App user's data as well as their friend's data. Although Facebook prohibited selling of user data collected via this method, it was sold and misused.

Medium in its case study has explained what and how psychology and behavior along with propaganda can manipulate people to do things intentionally or otherwise. They have linked it to the Cambridge Analytica Scandal by quoting how data driven marketing techniques can change behavior in target populations irrespective of the domain.

In[4] study has defined what the data actually looked like and explained the process of how it worked from

*Author for correspondence

Data Collection to Prediction. He has also mentioned the use of Machine Learning, role of AggregateIQ and OCEAN score.Thenextweb, in its article has explained how Cambridge Analytica made the use of profiling to swing neutral voters in favor of a particular candidate[5].

The Guardian noted that a scientific study conducted in 2013 revealed that liking curly fries related posts on Facebook gave clues about intelligence, similarly, liking hello kitty indicated political views [6-7].

3. Behind the Scenes

- 2010: Facebook launches Open Graph API for developers.
- 2011: Agreement with American FTC over consent for sharing user data.
- 2013: Cambridge Analytica founded as a subsidiary by SCL Group.
- 2013: A Cambridge Professor makes a personality detector quiz app for research purpose.
- 2014: The professor forms a company Global Science Research (GSR) and allegedly sells data to Cambridge Analytica.
- 2014 & beyond: Cambridge Analytica makes use of the data for various political campaigns.

4. Methodology

4.1 Data Collection/Mining

A Cambridge Professor created a Quiz App for Facebook for understanding user's psychology. The terms of service stated by Facebook at that time mentioned the developer was permitted to harvest user data for research purposes but did not explicitly mention if the developer was allowed to collect the user's friend's data as well.

Whenever a user would sign up to do a study, they would be given a survey to complete. The survey contained a Facebook Login Button using which the user logged into the app to do the survey.

As soon as the users logged into the App, they would have to authorize the App to have access to their user data. Although the authorization was only for their user data, inadvertently they authorized data collection of their friend's user data as well.

The data that was collected included their name, gender, location, ethnicity, education level, the pages they liked, the brand of clothing they wore, etc.

Starting with 250000 nodes, the professor/CA was able to collect data of approximately 75 million nodes.

4.2 Data Management (Categorization and User Profile Creation)

Based on the data that was collected, user profiles were created for each person and categorized accordingly.

Based on the geographic area they were residing in the US; user profiles were shaped.

E.g. Users near border areas were concerned with immigration, so they were clubbed in anti-immigration voter profiles. Users residing in hinterland were concerned with reduction in manufacturing jobs and construction of oil and gas pipelines through Native American villages, so they were clubbed together during profile creation. Ultra-High Net Worth Individuals residing in posh suburbs or downtown areas were concerned about tax breaks so that was the key factor for their profile creation.

Based on the brand of clothing users wore, their presumed political affiliations were derived.

E.g. Denim brands like Wrangler and L.L. Bean have been historically associated with conservative voters while Kenzo, another denim brand, is associated with liberal voters.Therefore, the user profiles were further refined according to their thoughts i.e. conservative or liberal. See the Figure 1.

Classification methods were used for categorizing user profile based on OCEAN scores. OCEAN scores stands for degree of Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism of the user see the Figures 2-4.

primary_Voters_CountyVoterID
primary_Voters_StateVoterID
primary_Voters_FirstName
primary_Voters_MiddleName
primary_Voters_LastName
primary_Voters_Gender
primary_Voters_NameSuffix
primary_Voters_BirthDate
primary_Voters_OfficialRegDate
primary_Voters_CalculatedRegDate
primary_Voters_BirthDay
primary_Voters_BirthMonth
primary_Voters_BirthYear
secondary_Voters_StateVoterID
secondary_Voters_FirstName
secondary_Voters_MiddleName
secondary_Voters_LastName
secondary_Voters_Gender
secondary_Voters_NameSuffix
secondary_Voters_BirthDate
secondary_Voters_OfficialRegDate
secondary_Voters_CalculatedRegDate
secondary_Voters_BirthDay
secondary_Voters_BirthMonth
secondary_Voters_BirthYear
primary_residence_AddressLine
primary_residence_ExtraAddressLine
primary_residence_City
primary_residence_State
primary_residence_Zip
primary_residence_ZipPlus4
primary_residence_HouseNumber
primary_residence_PrefixDirection
primary_residence_StreetName
primary_residence_Designator
primary_residence_SuffixDirection
primary_residence_ApartmentNum
primary_residence_ApartmentType
primary_residence_CensusTract
primary_residence_CensusBlockGroup
primary_residence_CensusBlock
primary_residence_UrbanRural
primary_residence_US_Congressional_District

primary_residence_US_Congressional_District
primary_residence_State_Senate_District
primary_residence_State_House_District
primary_residence_State_Legislative_District
primary_residence_Designated_Market_Area_DMA
primary_residence_County
primary_residence_Voters_FIPS
primary_residence_Borough
primary_residence_Village
primary_residence_Township
primary_residence_Town_District
primary_residence_Precinct
primary_mailing_AddressLine
primary_mailing_ExtraAddressLine
primary_mailing_City
primary_mailing_State
primary_mailing_Zip
primary_mailing_ZipPlus4
primary_mailing_HouseNumber
primary_mailing_PrefixDirection
primary_mailing_StreetName
primary_mailing_Designator
primary_mailing_SuffixDirection
primary_mailing_ApartmentNum
primary_mailing_ApartmentType
primary_mailing_CensusTract
primary_mailing_CensusBlockGroup
primary_mailing_CensusBlock
primary_mailing_UrbanRural
primary_mailing_US_Congressional_District
primary_mailing_State_Senate_District
primary_mailing_State_House_District
primary_mailing_State_Legislative_District
primary_mailing_Designated_Market_Area_DMA
primary_mailing_County
primary_mailing_Voters_FIPS
primary_mailing_Borough
primary_mailing_Village
primary_mailing_Township
primary_mailing_Town_District
primary_mailing_Precinct
secondary_residence_AddressLine
secondary_residence_ExtraAddressLine

Figure 1. Snapshot of what data was collected.

ename	etype	edate	eresult
Primary_2013-09-10	Primary	2013-09-10	Democrat
General_2012-11-06	General	2012-11-06	
Primary_2012-06-26	Primary	2012-06-26	Democrat
General_2011-11-08	General	2011-11-08	
General_2010-11-02	General	2010-11-02	
Primary_2010-09-14	Primary	2010-09-14	Democrat
General_2009-11-03	General	2009-11-03	
Primary_2009-09-15	Primary	2009-09-15	Democrat
General_2008-11-04	General	2008-11-04	
Presidential_Primary_2008-02-05	PresidentialPrimary	2008-02-05	
General_2006-11-07	General	2006-11-07	
General_2005-11-08	General	2005-11-08	
General_2004-11-02	General	2004-11-02	
General_2002-11-05	General	2002-11-05	
General_2001-11-06	General	2001-11-06	
General_2000-11-07	General	2000-11-07	
General_2014-11-04	General	2014-11-04	
Primary_2014-09-09	Primary	2014-09-09	
General_2013-11-05	General	2013-11-05	
PRI_BLT_2014-09-09	GovernorPrimary	2014-09-09	Democrat
PRI_BLT_2013-09-10	GovernorPrimary	2013-09-10	Democrat
PRI_BLT_2012-06-26	GovernorPrimary	2012-06-26	Other
PRI_BLT_2010-09-14	GovernorPrimary	2010-09-14	Other
PRI_BLT_2009-09-15	GovernorPrimary	2009-09-15	Other
PRI_BLT_2008-02-05	GovernorPrimary	2008-02-05	Democrat

Figure 2. Example of voter profile.

National Debt Importance Rank [1-10]	10
Gun Rights Importance Rank [1-10]	3
Traditional Social and Moral Values Importance Rank [1-10]	9
Environment Importance Rank [1-10]	5
Education Importance Rank [1-10]	4
National Security Importance Rank [1-10]	7
Immigration Importance Rank [1-10]	8
Socially Progressive Civil Rights Importance Rank [1-10]	6
Jobs and Economy Importance Rank [1-10]	1
Healthcare Importance Rank [1-10]	2
Registered Partisanship	Very Unlikely Republican
Unregistered Partisanship	Very Unlikely Republican
2016 General Election Turnout Propensity	Very High

Figure 3. Example of voter issue model of Prof. David Carroll.

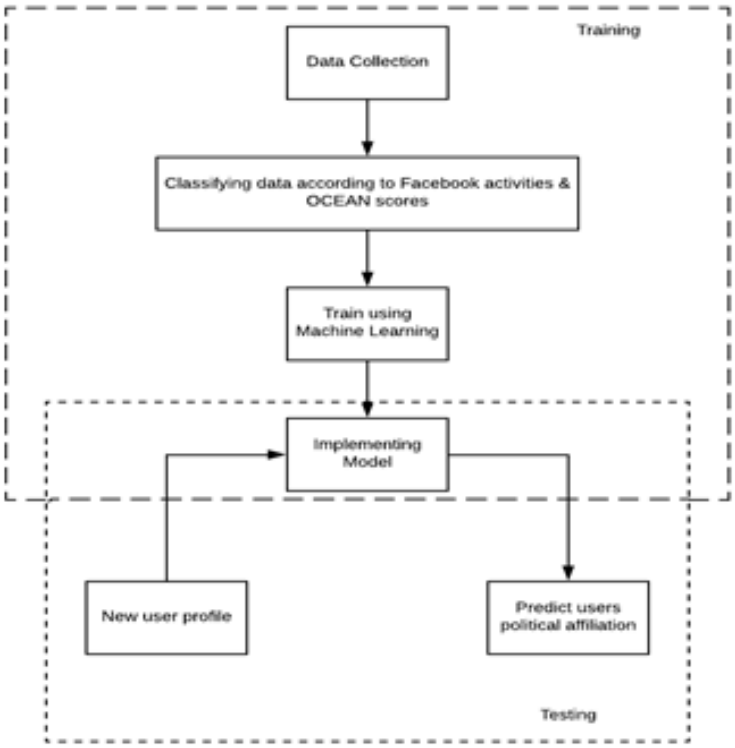

Figure 4. The entire process in block diagram.

4.3 Machine Learning

Regression used for training Machine Learning model on available data sets and predicting a new user's political affiliation.

User's Profile with psychographic data, their activities on Facebook (likes, posts, shares) and OCEAN score-all were used combinedly to predict information related to political affiliation. Machine Learning Model also considered the user's concern over some policy and the changes user wants to see the government to implement.

4.4 Micro-targeting

The result of the Model is used to create a valuable Campaign.

Users are presented with the advertisements keeping in mind their political views. Psychographic Micro-targeting Advertisements are presented in such a way that either it will enhance the belief of a user or it will reinforce the pre-conceived notion related to any party.

E.g. If a User is concerned with immigration policy (Anti-immigration User Profile Class) then he/she will be targeted with the ads explaining how a particular party is working in that area. Or, a rich class will be targeted with ads how well they are working on economic policies.

The ads are well customized to fit user's psychographics, aimed to yield maximum results.

5. Observations

Based on the above analysis, it can be observed that Social Media platforms such as Facebook and analysis companies such as Cambridge Analytica and unethical developers can use simple, trivial things such as posts, comments, likes and shares to gather sensitive information about user such as race, gender, orientation as well as provide a hint as to the political affiliation of a person. A few random likes can form basis for weirdly complex character assessment.

6. References

1. How Cambridge Analytic's Facebook micro targeting model really worked [internet]. https://www.iafrikan.com/2018/04/03/how-cambridge-analyticas-facebook-targeting-model-really-worked-according-to-the-person-who-built-it/. Date accessed: 03/04/2018.
2. C. Axford. Cambridge Analytic: A Case Study in Behaviorism Run Amok [internet]. https://medium.com/@craig.axford/cambridge-analytica-a-case-study-in-behaviorism-run-amok-de3d019e87c0. Date accessed: 27/03/2018.
3. B. Resnick. Cambridge Analytic's "psychographic micro targeting": what's bullshit and what's legit [internet]. https://www.vox.com/science-and-health/2018/3/23/17152564/cambridge-analytica-psychographic-microtargeting-what. Date accessed: 26/03/2018.
4. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. In Computer. 2018; 51(8): 56–59. https://doi.org/10.1109/MC.2018.3191268.
5. Clark B. How Cambridge Analytica leveraged fashion to sway undecided voters [internet]. https://thenextweb.com/plugged/2019/08/25/what-hell-telephoto-lens-phone-picture-photography/. Date accessed: 2006.
6. Cadwalladr C, Graham-Harrison E. How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool [internet]. https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm. Date accessed: 17/03/2017.
7. Just got my data from Cambridge Analytica/SCL by request. Yes, they do have correct voter and personal information about me [internet]. https://twitter.com/profcarroll/status/846392529838882816. Date accessed: 27/03/2017.