Prediction of psychological traits based on Big Data classification of associated social media footprints

Gagan Arora
Indiana University
2709 E 10th St
Bloomington, Indiana 47401
gkarora@iu.edu

ABSTRACT

Discusses the importance of digital footprints in evaluating person's psychological traits. We also reviewed few researches and articles which conducted studies in this field. We presented an algorithm at very high level of abstraction to understand how digital qualitative data can be translated to quantitative data to arrive at psychological traits. We concluded by providing few real life examples such as how Facebook likes can be used to evaluate psychological traits, how this research was used in last year elections and etc.

KEYWORDS

Big Data, Edge Computing i523, psychological traits, Big Data, Facebook Data, Social media, digital foot prints, five factor model, personality traits, elections, Facebook likes, Facebook comments, Instagram

1 INTRODUCTION

With the advancement of digital media and social media networks, there has been enormous amount of human activities, which is recorded as the digital footprints. According to IBM, in 2012 on an average 500 MB of personal data is uploaded to the online digital database daily. This data is either in the form of social media activities such as Facebook likes, Facebook comments, profile picture upload, tweets or in the form offline transactions where person goes to grocery shopping and pays using credit card. According to [6] China is investing heavy technological resources to mine this data along with person's financial transactions to build social credit system. This project is expected to be implemented by 2020. There has been studies [1] — [12], which analyzed the behavior outcomes of the digital profile with the actual characteristics of an individual. Interesting thing about these studies is that human behavior can be mapped statistically to define similarities and differences between individuals. This can further be used to build recommendation based system to enrich social medial networks such as Facebook, LinkedIn, and Twitter etc. These studies [1] to [12] further contributes in radically improving our behavior understanding of humans. [8] discusses about the predictability of individualfis psychological traits using statistical approach to arrive at the personality traits with certain confidence level. Psychological traits automation can further be used to enrich the quality of recommendation based systems and online search engines. [3] suggest how these studies [1] and [12] can be used to improve online marketing systems. With so many advantages on one side, on other side it possesses biggest challenge to the Data privacy [2] and [10]. Reason why these studies [1] and [12] provide better estimate of human psychological traits as compared to results of psychometric test because these study results [1] and [12] takes the data of prolonged history. However, psychometric tests on the other hands is for few minutes or hours where human can manipulate response in order to achieve desire results. Thus, these studies [1] and [12] can also be leveraged in employee hiring process where many companies still relies on psychometric tests.

2 DATA SOURCE OF BIG DATA IN DIGITAL WORLD

This section discusses how we can import, store and preprocess digital big data. This data can be fetched online via REST api or its direct available to download from website such as mypersonality.org. This site stores the social media data of close to six million participants. There are other sites like Stanford network analysis project, which contains enormous amount of data in the form of product reviews, Tweets, and social media data. Social medial sites like Instagram and Twitter provides public rest APIs through which we can access data, which is public. Other example is Amazon.com, which provides elegant web services to access product reviews. For preprocessing of this data, Python provides excellent libraries to access [via web service call] and preprocess data.

3 HUMAN BEHAVIOR AND PERSONALITY

[11] talks about various models, which can be used to describe human personality. Among all, five factor model [FFM] is proved to be the best model to describe human behavior, psychological traits and preferences: Openness, Conscientiousness, Extroversion, Agreeableness and Emotional stability. We have data, we have psychological traits, and biggest challenge lies in extracting value out of big data and mapping the result to psychological traits. To accomplish this challenge we can perform singular value decomposition to map the qualitative data to quantitative data. To elaborate this further let us take an example: we have a Facebook likes of 10 million people and we filter down top 100 Facebook pages, which are of relevance. Top 100 relevant pages are those, which can predict factors mentioned in FFM. Now we will prepare Boolean matrix with Facebook user profile on vertical axis and Facebook page as horizontal axis. In simple words row represents Facebook user and column represents Facebook page. We will mark the coordinate as one if corresponding Facebook user [on vertical axis] likes a page [on horizontal axis] otherwise zero. Therefore, matrix will look like this:

$$\begin{array}{c} \\ user_1 \\ user_2 \\ user_3 \\ user_n \end{array} \begin{array}{c} \begin{array}{cccc} fbPage_1 & fbPage_2 & \dots & fbPage_n \end{array} \\ \begin{pmatrix} 1 & 0 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix} \end{array}$$

These 100 Facebook pages is clustered, based on the five factors mentioned in FFM. First twenty pages will represent first factor, second twenty pages will represent second factor and so on. Next step would be to build correlation matrix that represents how each person is correlated with each other based on the five factors. This matrix will be N by N where is N is number of Facebook users in this experiment. This matrix will help to determine how similar Facebook users are. Which will help us to build the recommendation based systems because similar peoples tends to like same pages and share same psychological traits. This correlation matrix will look like this:

$$\begin{array}{c} \\ user_1 \\ user_2 \\ user_3 \\ user_n \end{array} \begin{array}{c} \begin{array}{cccc} user_1 & user_2 & \dots & user_n \end{array} \\ \begin{pmatrix} 1 & .75 & \dots & .85 \\ .75 & 1 & \dots & .91 \\ \vdots & \vdots & \ddots & \vdots \\ .85 & .91 & \dots & 1 \end{pmatrix} \end{array}$$

Step 1: Build binary matrix with Facebook user profile on vertical axis and Facebook page as horizontal axis.

Step 2: Populate the binary matrix with one and zero depending on if person has liked the page or not.

Step 3: Sort Facebook page columns depending on the factors mentioned in FFM.

Step 4: Use this matrix to build correlational matrix represents how each person is correlated with each other based on the five factors.

Step 5: Apply k mean algorithm to group Facebook users of similar factors mentioned in FFM.

4 COMPUTER BASED PERSONALITY JUDGMENT AND HUMAN BASED PERSONALITY JUDGMENT

Research [14] has shown computer based personality judgments are more accurate than those made by humans. According to [14] perceiving and judging peoplefis personality is an important component of living society. Many cognitive decision made by humans are based on the judgment they have in their mind. This research [14] has shown how advance machine learning algorithms and statistical tools can be used to predict the personality traits and compared the results with the human judgments. This research also addresses the issue of substantiating the qualitative aspects of behavior with the quantitative parameters. Computer based personality judgment is not only based on machine learning or statistics but computer vision algorithms can also be used to distinguish facial emotions and concluding psychological traits.

5 SOCIAL NETWORK AS A PERSONALITY TRAIT PREDICTOR

[9] studies suggest how valuable social network is in predicting the psychological traits. According to [9], It is considered as one of the valuable digital footprints to predict intimate personal traits. For instance, number of friends and their location can be used to grade first factor of FFM, which is openness. Person romantic partner can be detected depending on the social network overlap of each friend, which can further be analyzed to predict one's sexual preference. These predictions can further be statistically analyzed to [14] to know how accurate predictions are. We can use social network data on the algorithm discussed in the "Human Behavior and Personality" and conclude a very strong predictions on the psychological traits of a person. It has been in the news that 2016 elections were strategized with the help of the social media big data which will be discussed in the next section.

6 SOCIAL MEDIA BIG DATA AND ITS IMPACT ON POLITICAL ELECTIONS

[13] suggests how last year elections were revolutionized by the impact of big data of social media. Using statistical and machine learning algorithms on social media big data, political parties filtered down the data to identify their likely supporters and then channelized their strategy to win their votes. These strategies were less expensive than conducting campaigns at various places. Traditional analysis is generally based on the survey which is in the sense is limited [7] but now with the ease of big social media data, analysis is more accurate and conclusive. There has been sophisticated tools available that can predict the personfis race depending on his or her name and location. In recent election, political parties also combined social media data and public data [from census Bureau] to run sophisticated machine learning algorithm to pinpoint their supports. All these mentioned ways helped the political parties to micro target their supporters and gained their votes.

7 SOCIAL ACTIVITY, THE PREDICTOR OF PERSONALITY

[9] suggests Facebook profile of a user is not static rather it also contains enriched records of digital footprints such as likes, comments, reactions to other posts. Such activities materializes the connections between user and content. This information along with the other activities such as playlist, browsing logs, online shopping activities and google queries can be used to develop sophisticated highly predictive FFM set for a user and with a very high confidence level can predict userfis age, gender, intelligence religious view and sexual orientation [9]. Very interesting example from the [9] suggests "Users who liked Hello Kitty brand tended to have high openness, low conscientiousness, and low agreeableness" - strange but very interesting! [9] research further elaborate the importance of comments. Semantic analysis on comment can be analyzed to infer one's personality as shown by the research: [5] and [4].

8 CONCLUSION

We discussed various ways in which social medial data can be utilized to build five factor personality model for a user. Main

purpose here is to review the literature work done in this field and also presented the algorithm which can be used to translate qualitative data to quantitative data and how value can be extracted to build FFM for a user. We discussed computer based personality judgments are better than the human based personality judgments. We also touched based where social network can be used to predict userfis personality. As discussed earlier, these researches [1] - [12] have proved to impact the general election last year in United states. Finally we concluded by showing evidences how social activity can be used to build the FFM for a user.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and all the TA's for their support and suggestions to write this paper.

REFERENCES

[1] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. The Social fMRI: Measuring, Understanding, and Designing Social Mechanisms in the Real World. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 445–454. https://doi.org/10.1145/2030112.2030171

[2] Declan Butler. 2007. Data sharing threatens privacy. *NCBI* 449 (11 2007), 644–5.

[3] Ye Chen, Dmitry Pavlov, and John F. Canny. 2009. Large-scale Behavioral Targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, New York, NY, USA, 209–218. https://doi.org/10.1145/1557019.1557048

[4] Adam D. I. Kramer and Kerry Rodden. 2008. Word usage and posting behaviors: Modeling blogs with unobtrusive data collection methods. (01 2008), 1125-1128 pages.

[5] Samuel Gosling, Sam Gaddis, and Simine Vazire. 2007. Personality Impressions Based on Facebook Profiles. *ICWSM* 7 (Jan. 2007), 1–4.

[6] Lucy Hornby. 2017. China changes tack on fisocial creditfi scheme plan. eNewsPaper. (July 2017). https://www.ft.com/content/f772a9ce-60c4-11e7-91a7-502f7ee26895 China changes tack on fisocial creditfi scheme plan.

[7] Sean Illing. 2017. A political scientist explains how big data is transforming politics. vox. (March 2017). https://www.vox.com/conversations/2017/3/16/14935336/big-data-politics-donald-trump-2016-elections-polarization

[8] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805. https://doi.org/10.1073/pnas.1218772110 arXiv:http://www.pnas.org/content/110/15/5802.full.pdf

[9] Renaud Lambiotte and Michal Kosinski. 2014. Tracking the Digital Footprints of Personality. *IEEE* 102 (12 2014), 1934–1939.

[10] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. (06 2008), 111-125 pages.

[11] Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist* 48 (02 1993), 26–34.

[12] Kern ML Dziurzynski L Ramones SM Agrawal M Shah A Kosinski M Stillwell D Seligman ME Ungar LH. Schwartz H, Eichstaedt JC. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. (2013). https://www.ncbi.nlm.nih.gov/pubmed/24086296

[13] Chuck Todd and Carrie Dann. 2017. How Big Data Broke American Politics. eNewsPaper. (March 2017). https://www.nbcnews.com/politics/elections/how-big-data-broke-american-politics-n732901 How Big Data Broke American Politics.

[14] Youyou Wu, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *PNAS* 112 (01 2015), 1–5.

We include an appendix with common issues that we see when students submit papers. One particular important issue is not to use the underscore in bibtex labels. Sharelatex allows this, but the proceedings script we have does not allow this.

When you submit the paper you need to address each of the items in the issues.tex file and verify that you have done them. Please do this only at the end once you have finished writing the paper. To d this cange TODO with DONE. However if you check something on with DONE, but we find you actually have not executed it correcty, you will receive point deductions. Thus it is important to do this correctly and not just 5 minutes before the deadline. It is better to do a late submission than doing the check in haste.

A ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

A.1 Assignment Submission Issues

Do not make changes to your paper during grading, when your repository should be frozen.

A.2 Uncaught Bibliography Errors

Missing bibliography file generated by JabRef

Bibtex labels cannot have any spaces, _ or & in it

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

A.3 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues

A.4 Writing Errors

Errors in title, e.g. capitalization

Spelling errors

Are you using *a* and *the* properly?

Do not use phrases such as *shown in the Figure below.* Instead, use *as shown in Figure 3*, when referring to the 3rd figure

Do not use the word *I* instead use *we* even if you are the sole author

Do not use the phrase *In this paper/report we show* instead use *We show.* It is not important if this is a paper or a report and does not need to be mentioned

If you want to say *and* do not use *&* but use the word *and*

Use a space after . , :

When using a section command, the section title is not written in all-caps as format does this for you

`\section{Introduction}` and NOT `\section{INTRODUCTION}`

A.5 Citation Issues and Plagiarism

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material

A.6 Character Errors

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”

To emphasize a word, use *emphasize* and not “quote”

When using the characters & # % _ put a backslash before them so that they show up correctly

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

If you see a ffigure and not a figure in text you copied from a text that has the fi combined as a single character

A.7 Structural Issues

Acknowledgement section missing

Incorrect README file

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

Do not artificially inffate your paper if you are below the page limit

A.8 Details about the Figures and Tables

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

Do use *label* and *ref* to automatically create figure numbers

Wrong placement of figure caption. They should be on the bottom of the figure

Wrong placement of table caption. They should be on the top of the table

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, .jpg

Do not submit eps images. Instead, convert them to PDF

The image files must be in a single directory named ”images”

In case there is a powerpoint in the submission, the image must be exported as PDF

Make the figures large enough so we can read the details. If needed make the figure over two columns

Do not worry about the figure placement if they are at a different location than you think. Figures are allowed to ffoat. For this class, you should place all figures at the end of the report.

In case you copied a figure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

Remove any figure that is not referred to explicitly in the text (As shown in Figure ..)

Do not use textwidth as a parameter for includegraphics

Figures should be reasonably sized and often you just need to add columnwidth

e.g.

```
/includegraphics[width=\columnwidth]{images/myimage.pdf}
```

re