INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET) E-ISSN: 2395-0056
VOLUME: 07, SPECIAL ISSUE | JUNE 2020 WWW.IRJET.NET P-ISSN: 2395-0072
International Conference on Recent Trends in Science & Technology-2020 (ICRTST - 2020)
Organised by: ATME College of Engineering, Mysuru, INDIA

Analyzing Digital Footprints for Predicting Personality Traits of Social Media Users

Sameeksha Tandon[1], Manish Ahuja[2]

[1]Computer Science and Engineering Department, Institute of Engineering and Technology, Lucknow, India
[2]UCD School of Computer Science, University College Dublin, Dublin, Ireland

---

***Abstract*—**With evolving technologies over the past years, it has become possible to collect data about what people do online: what they browse, what they post on their social networking profiles, what they follow, their comments, their photos, and even their friends. All these activities can be aggregated to determine personality profile of a person. This personality profile can be used for: targeting advertisements to the user, presenting evidence about the user in legal proceedings, analyzing the personality of a candidate for their recruitment, etc. In this paper, we first present a review of several previous research works, which had attempted to perform various tasks based on a database of online textual content produced by people from particular regions. We have then demonstrated our methodology of using the tweets made by a person in order to identify their possible personality traits. This technique has extensive application in the field of recruitment. Using this system, recruiters can judge the personality of a candidate (negative, introvert, joyful, etc.), given a sufficient number of their tweets, to be able to fully understand what the candidate will bring to their company.

Keywords- Emotion classification; Logistic Regression; Naïve Bayes Classifier; Natural Language Processing; Personality analysis; Sentiment analysis.

1. INTRODUCTION

On a daily basis, each one of us contributes to a larger portrait of our online presence by leaving behind the traces of our online interests and activities. This portrait lets several companies in targeting potential users for a broad spectrum like marketing, fraud analysis, risk analysis, evidence acquisition, preservation of evidence for future use etc. The rapidly changing and increasing data across the world, calls for revisiting the already existing data collection and analysis methods. These procedures must evolve to maintain stability between digital and legal standards. Several methods and models have been proposed and designed to collect and analyze the footprints of users in the digital world. Understanding and examining low-level data to form high-level pieces of information is a very exhaustive process. Methods that are more responsive to the large volume of data should be chosen to guarantee reliability and accuracy. Today, psychological tests are conducted in almost every company to identify the capabilities and characteristics of candidates and for making a judgement about the suitability of candidates. This research work aims at analyzing public data of users which may be provided to recruiters to help them with their recruitment process. The public data may be collected from various networking sites like Twitter, Facebook, etc. This data collectively forms digital footprints of users. Digital footprints may portray either positive or negative image of users depending upon the type of activities users do online. This work aims at recognizing user personality traits which may be of help for the recruiters to let them decide the best fit for their company. In this paper, the personality traits data set [16] that has been used is divided into eight different categories in four pairwise combinations like:

- Introversion (I) – Extroversion (E)
- Intuition (N) – Sensing (S)
- Thinking (T) – Feeling (F)
- Judging (J) – Perceiving (P)

Also, for each of the above-mentioned category, a different label has been used to predict the most accurate personality trait. For instance, a user may belong to introversion, intuition and judging category at the same time. In this paper, state-of-the-art prediction technique has been used to closely examine the behavioral aspects of different users based on their tweets, comments, reactions on Twitter.

2. RELATED WORK

This section is divided into three sub-sections. In section II-A, literature survey related to analyzing personalities of USA people based on their Twitter profiles has been presented. Section II-B presents works related to disambiguating different online profiles belonging to the same person. While section II- C focuses on work that

examined different behavior of the same user on different social media platforms to find the most accurate personality of a user.

a.

Shafaan et al. [6] proposed the approach of predicting the personalities of people from USA using their Twitter profiles. The people were divided into five categories: general, political, sports person, bodies in the Business world, and those of Hollywood fame. The researchers have used previous month's tweets, profile data, facial features as shown in user's display profile on Twitter to examine and predict personality traits. All this data was stored in MySQL. In this paper, models with five different factors have been used to determine personalities of users which includes: openness, neuroticism, conscientiousness, agreeableness and extraversion. The authors collected profile images of over 54,784 users active on Twitter including their tweets from the previous month using Tweepy and Twitter search API. After collecting data, the researchers did Text analysis: analyzing age, sex, occupation, location, mental stability of users and then extracted images which includes: extraction of color, facial features, type, etc. After all the steps mentioned above, authors validated their work against the five-factor model. The authors have used Elastic net regularization and Linear Regression to analyze personality from Twitter profile image. Their work concludes that different personalities have different approaches in selecting profile images for their social media accounts.

b.

Anshu et al. [2] proposed a technique to disambiguate social media profiles from different platforms belonging to the same user. They mainly used Names and UserID for this purpose. The authors were able to achieve an accuracy of 99%. The authors collected and combined the information they extracted using Profilactic, FriendFeed, Social Graph API. Their main focus was the unification of multiple online profiles of the same user, as so many profiles may lead to security threats such as: profile cloning, profile theft that might lead to phishing, compromised profiles, stalking, profiling on online platforms by attackers and advertisers etc. The authors had used digital footprints from one social media account and along with that, they used automated tools to identify the same user on other social media platforms.

c.

Hasan et al. [3] in 2015 proposed a method to identify whether users portray analogous identities on different social media platforms. They have focused on two platforms: Twitter and Disqus. They have extensively studied the impact of Twitter and Disqus on the mindset and thinking of its users as they share their ideas, comments on posts, their sentiments etc. The authors have tried to identify and compare 105 people's Twitter and Disqus usage to find their interests, sensitivity towards a particular situation and to show some lexical differences and similarities related to Big5 personality traits [4]. Based on their research, the authors concluded that analysis of online presence on more than one social media platforms is needed for accurately examining the personality traits of users.

3. METHODOLOGY

A digital footprint leads to all the traces that we leave behind while surfing the web, whether the personal data issued deliberately by the user such as: on social media platforms (active digital footprints) or the information on which the control of user is limited such as cookie information or the IP address (passive digital footprints). In the modern era, the proportion of human actions such as social interactions have become facilitated by digital services. The increasing access to digital media enables online projects aimed at collecting profiles data and exploring the risks associated with them.

a. Sentiment Analysis

Sentiment Analysis is a field within Natural Language Processing (NLP) that develops systems that try to identify and get opinions within the text. The sentiment is a vital type of information expressed in human languages. Besides, these systems extract attributes of the expression like:

- **Polarity**: whether the speaker expresses a positive or negative opinion
- **Subject**: the thing that is being talked about
- **Opinion holder**: the person or entity that expresses the opinion [13]. Using this approach, we have tried to find out the sentiments of different people in different scenarios.

Description of dataset*:* The dataset consists of Facebook status messages which may be either photo, video or links as shown in fig. 1. It contains 3220 rows in CSV format, with the following columns: [11]

- status_id - unique status ID generated by Facebook
- status_message – timeline Photo/video or linked page's title otherwise
- link_name - text written by in the link
- status_type – photo/video or link
- status_link -status link
- status_published - timestamp
- num_reactions - number of reactions
- num_comments -number of comments
- num_shares -number of shares
- num_likes - number of likes
- num_loves - number of love reactions
- num_wows - number of wow reacts
- num_hahas - numer of haha reacts
- num_sads – number of sad reacts
- num_angrys – number of angry reacts

Implementation*:* Initially, we have read the csv file and pre-processed the data to filter the text required for analysis. The pre-processing involves:

- Removing all emoji symbols from the text using a RegEx.
- Removing all extraneous separator characters

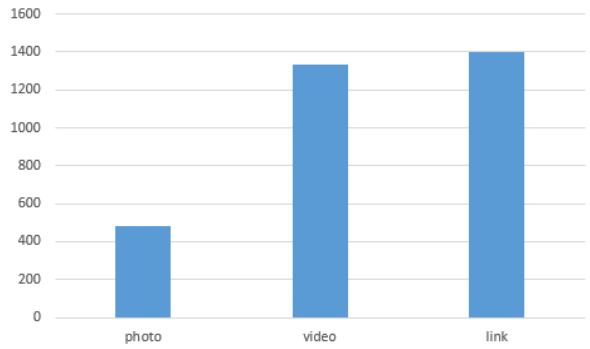

Figure 1. Distribution of photo, video and links

Then we have used Natural Language Toolkit Sentiment Intensity Analyzer Python package. After this, we have used the polarity_scores method with the argument "compound" so that we can get a polarity score. Positive polarity is indicated by a score of +0.5 or more, negative polarity is indicated by a score of -0.5 or below, and neutral polarity score lies in between.

A drawback of this method is that although we can classify a text either into positive, negative or neutral class, there may be several possible emotions attached to the sentences simultaneously. Not all emotions can be covered under the blanket in terms of positive, negative or neutral. Hence, we have then applied the Emotion Classification to predict the emotions along with the sentiments.

b. Emotion classification

Sentiment analysis can help us in identifying the general feelings of candidates like: negative, neutral, positive. Thus, analyzing sentiments can only help us in giving a general opinion about candidates with respect to certain conditions. However, there are situations where only the above-mentioned sentiments are not sufficient to draw adequate conclusions regarding candidates. Therefore, after Sentiment analysis, we have done Emotion analysis to identify various emotions of a candidate like: excited, angry, bored, sensitive, etc. Today, Emotion recognition is a widely used feature based on techniques

of Artificial Intelligence. On social networking sites, people make use of emojis in their text to express their feelings on different subjects. Emotion analysis can deal with text, audio, video, image etc. Emotion analysis is basically used to get a vivid view of the emotions of candidates on particular articulate matter.

Description of dataset: This dataset is of 'human emotions' which is labelled in 7 different categories, and has 2 columns and 7652 rows [12]. Fig. 2 shows the percentage of different emotions with respect to total number of samples. The first column is a string of different emotions having one of the following values:

- Fear
- Anger
- Sadness
- Disgust
- Shame
- Guilty
- Joy

. The second column is a string of user's comments.

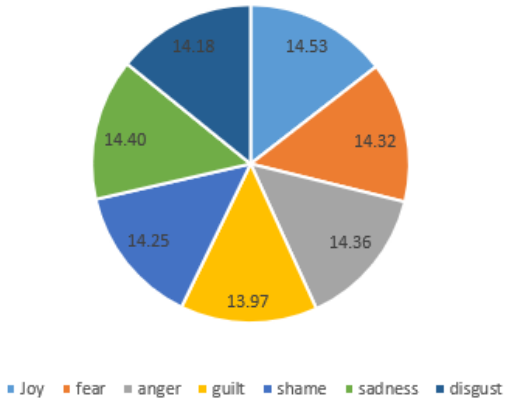

Figure 2. Percentage of emotions with respect to total number of samples

Implementation: The first 6000 rows have been selected for training data. Further, pre-processing has been done to get the final training set. The Naive-Bayes classifier has been used for the classification. Naïve-Bayes algorithm is a kind of probabilistic algorithm that uses Bayes' theorem to predict the category of text. It is effective and widely used for the classification problem [14]. Naïve-Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature. Bayes' theorem is a way to figure out conditional probability. The conditional probability is the probability of an event happening, given that it has some relation to one or more other events.

$$P(A/B) = \frac{P(B/A) \cdot P(B)}{P(A)} \quad (1)$$

For e.g.: the probability of getting a parking space is connected to the time of the day we park and the location where we park. Once the model is trained using Naïve-Bayes' classifier, with all parameters as default value as in textblob.classifiers v0.15.2, we have tested the trained model to get the emotion out of the manually entered text as shown below:

nb.classify("Ireland has the perfect weather") output: 'joy'

Further, we have evaluated the model performance by calculating the accuracy of the model. For this model, it is coming as 55.8%. Although Naive Bayes algorithm is fast and simple to implement, its major drawback is that it requires the predictors to be independent of each other. However, it is not always the case as the predictors are mostly dependent on each other. To remove this problem, we have used Logistic Regression algorithm in the next section.

c. Personality Analysis

Extending the work done in Emotion analysis, several comments by the same user can be aggregated in order to determine the personality of the user. The recruitment process in today's world is not the same as it was few years back. Companies don't want employees with just the technical skills set, they want employees who have a positive mindset which would help the company grow. Now, companies analyze the candidates' social networking profiles to get a better understanding of their lifestyle and nature. There is no need of traditional psychometric tests anymore. With the model implemented in this work, companies would be able to analyze the overall nature of their future candidates using their posts on various social networking platforms. This can help the employers in selecting the right candidate with the right attitude towards work.

Dataset: The first column is a string of four capital letters (one from each indicator) based on Meyer Briggs personality indicators [16], which are:

1) Introversion (I) – Extroversion (E)
2) Intuition (N) – Sensing (S)
3) Thinking (T) – Feeling (F)
4) Judging (J) – Perceiving (P)

Consider, for instance, a person is an extrovert and prefers intuition, thinking and judgement would be labelled as: ENTJ in this system. Similarly, a total of 16 distinct combinations across the above 4 axes are possible. The second column is a very large string of past fifty tweets made by a user. Each tweet is separated by triple pipe symbols (|||). A minimum of 50 tweets per user has been considered in this paper for improving the accuracy of the model. Fig. 3 shows the distribution of different personality traits.

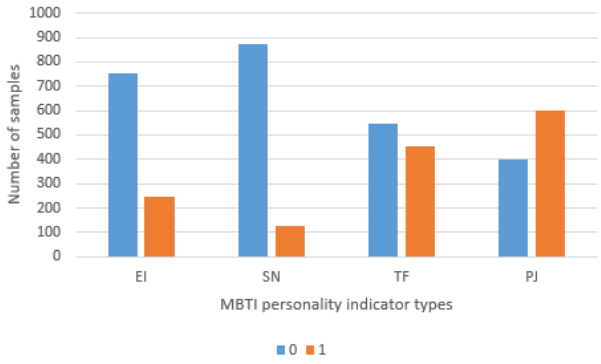

Figure 3. Distribution of personality traits

Implementation*:* It is a supervised learning method which uses the labelled dataset to train the model. This model can further be used to test different combinations of input values, and for classification prediction in such a way that output takes values for the given set of input features. These input features basically act as independent variables or predictors and the target is a dependent attribute. Regression identifies the relationship between predictor and target. Logistic Regression was chosen, keeping in mind the following points:

- It produces a more informative output as compared to other classification algorithms.
- It lets us unveil the hidden relationship between the variables.
- It is a very efficient approach as it does not require many resources for computation.
- It can be easily regularized and can be quickly implemented.

As our work deals with the fitness of a candidate in a workplace i.e. either he or she can be declared as fit or not fit for the employer (which is a linear problem), Logistic Regression can efficiently implement this idea by

categorizing the candidates into either of the two categories: 'fit' or 'not fit' [15].

We have first processed the data to count the number of times a person's tweet contains dots, exclamation marks, etc. in their tweets. This results in the addition of seven new numeric columns. we have then used a Newton-CG (Newton Conjugate gradient) Logistic Regression model on these columns with at max 100 iterations, the other parameters being default values in the Python 3 package scikit-learn 0.22. The code runs with one type indicator at a time. Our code produces an accuracy of 75.6% when running on the test dataset with type indicator - extroversion or introversion. The test dataset consists of 1000 randomly selected rows from the dataset.

4. CONCLUSION AND FUTURE WORK

This research work deals in with identifying the emotions of candidates and based upon those emotions, preference can be given to the candidates that fit for a particular role. For e.g. extrovert candidates are fit for roles such as marketing, teaching, etc. Thoughtful candidates are fit for product-based companies where candidates would come up with new innovative ideas. The first phase of this paper deals with sentiment analysis using Naïve Bayes classifier. This classifier performs only when predictors are independent of each other. However, real-world problems are such that elements are dependent on other factors. Therefore, Naïve Bayes is not a perfect solution to many problems. Also, in the first phase, our work dealt in with sentiment analysis, which only categorically classifies a person's sentiments and does not provide a clear understanding of personality traits. In addition to this, Sentiments can be identified only through textual data, whereas human interaction is not limited to text alone. Logistic regression was chosen in our research work because it allows us to get a confidence score for several classes of personalities. For each personality trait, we get a confidence value (between 0 and 1) representing how confident our neural network is in judging that personality trait. Naive Bayes classifier was chosen for similar reasons. The Naive Bayes classifier converges quickly than other discriminative classifier model because in this the assumption of conditional independence holds for a certain degree. Thus, we need comparatively less data for training our model. We have used Logistic regression and Naive Bayes to train our models. We could as well have used other more complicated models, like Support Vector Machine or a deep neural network, however, fine tuning them for our purposes could be extremely tricky. The presence of several hyper- parameters as well as complicated structure makes it extremely difficult.

In another direction, our work can further be extended by making predictions for several other more complicated personality traits. These can be agreeableness, openness, determination, calmness, etc. Such features would help employers better analyze their prospective candidate. However, it can be a challenge to accurately judge such complicated traits from just the text-based social media posts of a person. Therefore, another extension of this work can be to use audio or video data (gathered from social media) as well. However, both the collection and pre-processing of non-text data is a huge challenge in itself.

References

[1] Zedadra, O. Zedadra, and A. Kebabi, "Analyzing Traces in an Informal Social Environment" in 2019 International Conference on Networking and Advanced Systems (ICNAS), Annaba, Algeria, June 26-27, 2019.

[2] A. Malhotra, L. Totti, W. Meira, P. Kumaraguru, and V. Almeida, "Studying User Footprints in Different Online Social Networks," in ASONAM '12 Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), August 26-29, 2012, pp. 1065-1070

[3] H. A. Maruf, N. Meshkat, M. E. Ali, and J. Mahmud, "Human behaviour in different social medias: A case study of Twitter and Disqus," in 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, August, 25-28, 2015

[4] P. Rosen and D. Kluemper. The impact of the big five personality: Traits on the acceptance of social networking website. In AMCIS, 2008.

[5] W. Chung, S. He, D. D. Zeng and V. Benjamin, "Emotion Extraction and Entertainment in Social Media: The Case of U.S. Immigration and Border Security," in 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), Baltimore, MD, USA, May 27-29, 2015.

[6] S. K. Bhatti, A. Muneer, M. I. Lali, M. Gull, S. M. U. Din, "Personality Analysis of the USA people using Twitter profile pictures," in 2017 Inter- national Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, December, 30-31, 2017.

[7] DataCamp Community. (2019).Top 5 Python IDEs For Data Science.[online] Available at: https://www.datacamp.com/community/tutorials/data-science-python-ide [Accessed 12 May 2019].

[8] Pandas.pydata.org. (2019). Python Data Analysis Library — pandas: Python Data Analysis Library. [online] Available at: http://pandas.pydata.org/. [Accessed 18 May 2019].

[9] [online] Cs231n.github.io. (2019). Python Numpy Tutorial. Available at: http://cs231n.github.io/python-numpy-tutorial/.[Accessed 19 May 2019].

[10] [online] https://www.tutorialspoint.com/artificial_intelligence/artificialintelligence_natural_language_processing.htm [Accessed 25 May 2019]

[11] [online] GitHub (2019).minimaxir/interactive-facebook-reactions. Available at: https://github.com/minimaxir/interactive-facebook reactions/blob/master/data/bleacherreport_facebook_statuses.csv [Accessed 13 May 2019].

[12] [online] Available at: https://raw.githubusercontent.com/bogdanneacsa/tts-master/master/ISEAR/ise_processed[Accessed 14 May 2019].

[13] [online] Available at: https://monkeylearn.com/sentiment-analysis/ [Accessed 23 May 2019].

[14] [online] Naïve Bayes Classifier. Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Accessed 18 May 2019].

[15] [online] Available at: https://machinelearning-blog.com/2018/04/23/logistic- regression-101/ [Accessed 4 July 2019].

[16] [online] Available at: https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1 [Accessed 6 July 2019]

[17] [online]. Available: https://www.crunchbase.com/organization/continuum- analytics#section-overview. [Accessed 12 May 2019].