WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Digital Footprint in Social Media

[1]Sreyas Krishna S, [2]Dr. Febin Prakash

[1,2]Department of CS&IT, Jain (Deemed-To-Be_University), Bangalore, India jpc222705@jainuniversity.ac.in, febin.prakash@jainuniversity.ac.in
DOI: https://doi.org/10.55248/gengpi.5.0324.0646

ABSTRACT-

In today's world, social media use has grown progressively more widespread. The way we connect, communicate, and share information with others has been completely transformed by it. Furthermore, social media has enabled the development of online communities and e-participation in addition to bridging geographical divides. Social media has additionally shown itself to be a useful instrument for political campaigns and increasing public awareness of significant social concerns. Social media has given people and organizations a forum to voice their thoughts, exchange experiences, and have deep conversations. All things considered, social media's introduction has had a significant influence on communication and created new opportunities for involvement and engagement in both social and personal issues. In conclusion, social media has revolutionized communication by providing a platform for real-time engagement, breaking down geographical barriers, and promoting accountability.

Keywords—Natural Language Processing, Big Data, Social Media, Digital Footprint, Personality Prediction.

1. Introduction

Customers create and leave digital traces on digital platforms, known as big data digital footprints, when they use and interact with different media channels, such as social media. Researchers and practitioners are more interested in consumers' social media-led digital footprints due to the growing usage and involvement with social media. Digital footprints are defined as social data generated by consumers during their media channel interactions. These digital traces include behaviours, moments, memories, and identities. Social media companies that compile these massive digital histories have the ability to ascertain user behaviour and buying patterns on digital platforms.

The use of social media has increased dramatically and is now a necessary aspect of daily life for consumers. The creation of digital footprints has greatly risen with the introduction of Web 2.0. By 2020, 44 zettabytes of data would be generated, an estimated 44 times more than in 2009 **[5]**. Furthermore, there are a lot more social media touchpoints thanks to the rapid expansion of 4G networks, cloud computing, and mobile phone technology. Due to consumers' constant connectivity to smart devices (smartphones, tablets, wearables, Cortana, Siri, and Alexa, among others), service providers are left with enormous digital trails. Moreover, social media user-generated material will serve as the primary means of enhancing public administrative bodies' and private companies' knowledge bases. Based on various social media channels and vehicles, Tuton and Solomon separated the use of social media into four zones: social community engagement, social commerce, social publishing, and social entertainment.**[5]** Consumers share comments, images, videos, blogs, bookmarks, reviews, ratings, and social shopping on channels within each of these zones, creating their digital identity while also connecting with official applications, among other things. These digital trails, which are essentially mandated by companies, display their hobbies, social and cultural identities, and geographic and vocational affinities. Additionally, by utilizing advanced analytics to analyse client sentiments and content, businesses may better understand their behaviour and build more comprehensive profiles thanks to these digital trails.

Consumers may or may not be conscious of the digital footprints they leave for businesses like Google, Amazon, and Facebook, but they use social media excessively. Social media platforms (Facebook, Twitter, Instagram, and so on) and their providers have used personal data to follow users' behaviour through intrusive and widespread crawling, which has completely changed the ways in which businesses can create value. By making connections between data, drawing conclusions, and interpreting data, they employ algorithms to produce insightful information. In addition, businesses are competing more and more to find cutting-edge solutions to mine digital trails and outperform competitors. Therefore, large data digital footprints have enormous managerial significance because they can generate revenue and promote development.

2. Literature Study

Social media's pervasiveness has created a complex web of digital footprints that record our online activities, preferences, and habits in a way that is both distinct and dynamic. Researchers have been enticed by this data mine to extract insights into people's personalities, which are their fundamental characteristics. A recent study explores this intriguing area by determining whether personality traits can be predicted exclusively from social media traces. This study uses the rich data landscape of social media to explore several machine learning algorithms for personality prediction **[6]**. Its

fundamental tenet is that there are subtle indicators of our psychological tendencies hidden in our online conduct. The study intends to create a computational portrait of user personalities by means of a thorough investigation of language usage, social interactions, and content preferences.

2.1 Working of Social Media

It's critical to comprehend social media to solve the civil discovery issues that these platforms raise. A social media website is, in its most basic form, an internet platform for sharing content with an audience or interacting with people. While sharing information with others is at the core of this straightforward definition, social media platforms also provide users with a range of privacy settings that are meant to allow them some control over who can see their personal data.

Social media websites are also multifaceted platforms that motivate users to carry out additional online tasks, either through the social media site itself or through applications from third parties that are integrated with the social media site. As a result, layers of data are present on social networking websites that are either user-generated or have been assembled by the platform. Put differently, a social media account stores information about a user's activities, including purposefully shared content and other transactional data derived from their entire online activity. Facebook is the most popular social media platform. It boasted more than a billion monthly active users as of the end of 2012**[7]**.

Anyone can create a personal profile on Facebook, and users can choose the audience they want to share content with by adjusting various privacy settings. Since Facebook is currently the industry leader in social media, knowing more about its features can help you comprehend social media in general. **[7]**

2.2 Important Social Media Services and Capabilities

Social media platforms are about sharing. But Facebook has several facets, just like most social networking platforms. First, Facebook gives users the ability to send and receive private messages to other users, as well as to directly engage with them one-on-one or through Timeline postings that are visible to a wider audience.

Furthermore, Facebook allows users to do additional tasks on the internet by integrating its services with other websites or third-party applications. Consequently, Facebook's features and functionalities allow users to accomplish a wide range of jobs on the platform. Making connections and communicating with people is the primary purpose of social media.

Users of Facebook, for instance, have a Timeline that serves as their customized profile and is the primary page that visitors see when they visit a user's Facebook page. In general, the user can manage which kinds of information are accessible to them and what is shown on the Timeline. A user's profile and cover photo, as well as a chronological list of posts made by the user or others, are all included in the timeline information. More images, videos, audio files, and other multimedia files can be uploaded by the user and displayed on the Timeline. Additionally, the Timeline has an "About" part that includes personal information about the user, including a list of pages liked, persons the user "Follows," photographs or comments the user has "Tagged," and information about the user's activities on other Websites **[7]**.

In addition, a user's Timeline map, which shows a graphical representation of the places they have been, is updated whenever they "Check In" or add a location to a story on Facebook.

The relationship that is established when two people decide to become Facebook "Friends" frequently determines access to data on that person's Timeline. Users can usually adjust the amount of personal information that is exposed to the public, to Friends alone, or to specific subgroups of Friends using privacy settings. However, the private sections of a user's Timeline are typically visible to their "Friends" at the very least.

2.3 Connectivity with Additional Websites or Services

Many social media websites integrate with other websites or services in addition to promoting communication among users. Facebook, for instance, has developed "Social Plugins" that allow users to click "Like" buttons on websites other than Facebook in order to record their online activity. According to Facebook, social plugins enable users' "social experiences" on the internet more personalized by enabling them to share and view content from other Facebook users when they visit websites run by third parties. A user who is simultaneously connected into Facebook and visiting other websites can see the features and outcomes of these social plugins.

Facebook also incorporates a range of third-party applications, or "apps," into its websites. Examples of these include calendars, games, news services, music streaming services, restaurant and travel reviews, discount apps, and other social media websites. Certain third-party applications employ geotagging capabilities on the user's phone to locate position, or they post automatically to the user's Timeline. Users can share even more information about their interests and habits and do a variety of tasks without ever leaving Facebook by utilizing third-party apps.

When users visit Facebook, they also see "Sponsored Stories" and adverts. Customized and targeted advertisements based on user choices and likes are made possible by Facebook's advertising framework. Because Facebook gathers personal data, businesses might potentially identify their target audience more quickly.

Some examples of how social networking websites are multifaceted and combine their functioning with external services and features are social plugins, third-party apps, and targeted adverts. However, social media websites' various features also broaden the range of user data that is generated—and

frequently saved—in a social media account. As a result, Facebook has developed into a multipurpose platform even though its core purposes were sharing content with friends and managing a personal timeline.

2.4 Configuring Privacy and Storing Data

One important component of social networking networks is their privacy settings. However, a social media account's content as well as the people it is meant for are multifaceted and intricate. Social networking websites gather a lot of personal data about their users due to their extensive functionality. As a result, a significant quantity of personal data is nevertheless kept on file as part of the user's social media account, even though individuals have some discretion over who can view their content.

Most Facebook users are aware of and utilize some of the available privacy options. Facebook does not, however, guarantee user privacy, as stated in its terms of use, and it often updates these privacy settings.

Although a user's privacy settings have a significant impact on the content they voluntarily choose to share, Facebook also gathers and retains data about their account. Put another way, Facebook or other third-party apps generate or assemble some of the content that users do not knowingly submit. The user's account contains both personally identifiable information and Facebook-complied information. It's unclear exactly what information Facebook retains and for how long.

However, Facebook allows users to easily download a large amount of their own data directly from the platform. Therefore, the extent of personal data saved by the social media platform is a crucial element of a social media account, in addition to the availability of user-controlled privacy settings.

3. NLP Techniques

Defining intentions is the first step towards using various natural language processing techniques. This means that it was crucial to specify the chronology and interaction between the questions to create a coherent dialog and sequence. As a result, there were input and output contexts for each intent **[10]**.

Our digital footprints leave a quiet trail of thoughts, preferences, and interactions in the ever-expanding world of social media. These traces have the capacity to unveil the mysterious fabric of our characters, much like strewn parts of a puzzle. An intriguing route to realizing this promise is using Natural Language Processing (NLP), a potent instrument in the field of artificial intelligence.

Imagine exploring a sea of text data that includes tweets, status updates, comments, and more, each word whispering something about the person reading it. By removing meaning and revealing patterns that are hidden from view, NLP approaches help us navigate these enormous waters. Sentiment analysis allows us to explore the underlying emotions and spot happy, angry, or anxious expressions. By revealing the subjects that capture people's attention, topic modelling assists us in mapping out the terrain of interests. By examining language style, one can uncover hints regarding personality qualities. For example, the frequent use of first-person pronouns may indicate extroversion, while word choices that promote cooperation may indicate agreeableness.

However, the trip continues even after a word is understood. NLP enables us to understand the complex connections between words and reveal the deeper meaning weaved throughout online communication. We can learn more about social networks and how people connect and engage with each other in online groups by breaking down phrase structures and examining conversational patterns. When paired with unique personality insights, this web of relationships creates a more comprehensive picture that provides hints about an individual's social identity and online character.

It is not a simple task to train the data for this personality prediction. For the data to be relevant and reliable, it must be carefully chosen and cleaned. It is imperative to represent this textual data in an understandable manner for algorithms, which frequently calls for sophisticated methods like feature extraction and word embedding. When the data is prepared, machine learning algorithms are taught to find patterns and connections between linguistic features and personality traits. These algorithms are aided by natural language processing (NLP) techniques. Complex algorithms such as supervised learning or deep learning, which learn from labelled data to predict unknown data, are frequently used in this training process.

But there is pre-caution to be taken when stepping into this interesting world. Privacy is a major concern, and using personal data for personality prediction has ethical ramifications that should be carefully considered. To maintain fairness and prevent biased results, potential biases in the data or algorithms must be addressed. To ensure that these potent instruments are used sensibly and morally, transparency and accountability are crucial.

4. Methodology

The preparation of data and the creation and tracking of themes comprised the two primary components of the approach used in this investigation.

The data were gathered, filtered, refined, and pre-processed using natural language processing (NLP) techniques in the first phase. Following the completion of these procedures, the second section addressed "what people talk about" and described how topics are tracked and modelled. In a comparison methodology, two processes were evaluated: the creation and tracking of artificial topics, as well as the selection of top expert term sets.

The approach establishes a list of pertinent topics based on predetermined "expert terms" and models each topic independently over a few weeks. Using the specifications of the artificial topic, the second technique generates an artificial topic known as "expert terms" and models subsequent topics.

4.1 Topic Modelling

Following the NLP pipeline's preprocessing of all the posts, subjects were modelled using the "cluster able words" that were taken out of each post. The well-known "k-means" computational cluster formation approach, which requires the specification of the number of output clusters "k" in advance, was used to create the number of topics as input for the topic modelling procedure. A standardized technique must be used to determine the number "k" to build a consistent and replicable methodology that can be implemented in data sets of varying sizes and sources.

An average value was determined as the result of this analysis's series of correlations between various topic counts and the coherence assessments made in each of them.

Segmentation, probability estimation, a confirmatory measure, and the establishment of a final aggregation value were combined to quantify topic coherence. The average aggregation value for all themes, or the "coherence score," tended to exhibit a distinct peak on a curve that was described by the evaluation of coherence output.

4.2 Expert Terms

Expert terms were manually listed words from a semantic domain that were specialized to a given domain. Expert terms were utilized to assess the issues' relevance rather than being employed in the modelling process. We identified postings that may be associated with cultural events by looking for phrases like "culture, museum, or exhibition" that belong in the same semantic field. Following the identification of a "key week,"**[9]** the subjects that were created were examined closely. "Relevant topics" were those that had a greater ratio of expert terms per post. The ratio of expert terms per post for each week was examined independently, and the topics were grouped based on the highest to lowest value. This clustering process was used to estimate the number of relevant topics to take into consideration.

5. Challenges & Solutions

Researchers now face several challenges because of these new data, from questions about the authenticity of the data and how it was sampled to moral dilemmas with its application. Online data create a contradiction when it comes to privacy protection: although they are too disclosing when it comes to privacy protection, they are also not disclosing enough when it comes to giving social scientists the demographic background data they require. The comprehensive demographic pro-file information that is typical in survey research is frequently absent from online data. For instance, even though Twitter data are available to the public, many users only submit incomplete, fictitious, or confusing profile information, which makes it challenging for researchers to link the characteristics of network nodes or the content of tweets to fundamental demographic indicators.

Here are some possible answers to these problems:

To meet the objectives and demands of both the academic community and industry, new procedures and institutional frameworks are required. Online businesses fight hard to draw in academic talent, especially social scientists, and some of them have university relations sections that can support joint study. Additionally, to access and handle massive semi-structured data collections, advanced programming and other technical abilities are needed. Doing controlled studies, like the Facebook experiment, that vary exposure to a potential epidemic is one way to find a solution. When observational data is the only available and experimental methods are impractical, researchers can distinguish between influence and selection by comparing the presence of the contagion between egos that have been exposed to it and those that have not, using an instrumental variable that is linked to the alter's exposure to the contagion but not the ego's **[8]**.

6. Conclusion

Social networking platforms have ingrained themselves into our lives in today's technologically advanced world, leaving behind a wealth of data known as digital footprints. To explore the intriguing field of personality prediction based only on these imprints, this study delves into the fascinating field of Natural Language Processing (NLP) approaches. The results demonstrate how promising natural language processing (NLP) is for deriving insights about personality traits from language use, social interactions, and content preferences. Through the identification of minute hints concealed in online actions, we can initiate the creation of computational portraits that illuminate people's psychological tendencies. These kinds of findings have the power to transform several industries, including early mental health intervention, human-computer connection, and targeted marketing.

But entering this fascinating field requires thorough consideration of ethical issues. Maintaining individual rights requires ethical data practices, which are crucial in the ongoing fight against privacy violations. Vigilant attention is needed to avoid algorithmic bias and potential exploitation of personal data. Those using these potent tools must be transparent and accountable.

The conceptual validity of the data collection and cleaning techniques described in this paper was supported by the trends' consistency over all weeks and their resemblance to other activity studies from various disciplines, such as energy demand curves or traffic congestion curves. In keeping with Manovich's [9] strategy of using social media to depict "the everyday" rather than "the extraordinary" and placing online social material inside the framework of mass culture production, social media use was also linked to significant routine activity patterns.

In the end, the study of digital footprints and personality prediction is a multifaceted field full of obstacles as well as opportunities. Responsible innovation and moral considerations must direct our actions as we travel along this path. We can unlock the promise for a future where technology improves our lives without compromising our privacy or individuality by utilizing NLP while upholding individual rights.

References

[1] Deeva, I. (2019). Computational personality prediction based on digital footprint of a social media user. *Procedia computer science*, *156*, 185-193.

[2] Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in cognitive sciences*, *19*(12), 771-782.

[3] Getachew, A., & Beshah, T. (2019). The role of social media in citizen's political participation. In *ICT Unbounded, Social Impact of Bright ICT Adoption: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2019, Accra, Ghana, June 21–22, 2019, Proceedings* (pp. 487-496). Springer International Publishing.

[4] Toor, S. I. (2020). Social Media as a Mediator in Political Communication: A Literature Review to Explore its Effects on.

[5] Muhammad, S. S., Dey, B. L., & Weerakkody, V. (2018). Analysis of factors that influence customers' willingness to leave big data digital footprints on social media: A systematic review of literature. *Information Systems Frontiers*, *20*, 559-576.

[6] Valanarasu, R. (2021). Comparative analysis for personality prediction by digital footprints in social media. *Journal of Information Technology and Digital World*, *3*(2), 77-91.

[7] McPeak, A. A. (2013). The Facebook digital footprint: Paving fair and consistent pathways to civil discovery of social media data. *Wake Forest L. Rev.*, *48*, 887.

[8] Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*, 129-152.

[9] López Baeza, J., Bley, J., Hartkopf, K., Niggemann, M., Arias, J., & Wiedenhöfer, A. (2021). Evaluating cultural impact in discursive space through digital footprints. *Sustainability*, *13*(7), 4043.

[10] Mancheno Gutiérrez, M. F. (2021). Chatbots as educational assistants: teaching about the digital footprint.