International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 09, Issue: 06 (November-December, 2023)

Automated Derivation of Marketing Analytical Insights from Social Media Information

Ravi Kumar [1], Sarwesh Site [2]
[1,2] Department of Computer Science and Engineering
All Saints' College of Technology, Bhopal, India

Abstract

In contemporary digital marketing, effective audience segmentation is crucial at both aggregate and individual user levels. Personas, representing fictional user profiles, serve to humanize data, offering distinct identities to audience segments. Marketers utilize personas to gain deeper insights into their customers and make efficient marketing decisions for each segment. Developed within the realm of big data analytics and artificial intelligence, this article proposes a solution that not only demonstrates the viability of automatic persona generation from social media data but also illustrates the design and implementation of a highly scalable, expandable, and accurate system for this purpose. The proposed solution comprises various components, each assigned specific tasks, including data collection, enrichment, clustering, and persona generation. Multiple classifiers are employed to extract insights from the data, subsequently utilized to create customer segments. The integration of these components into a stream-processing architecture mitigates bottlenecks, ensuring a high level of separation of concerns. The final outcome reveals user segmentation with exceptional accuracy, precision, and recall measures, all exceeding 0.90%.
Keywords: Social Media, Marketing, Machine Learning, Federated Learning, Data Analytics

1 Introduction

Social media data analysis plays a pivotal role in shaping effective digital marketing strategies. This process involves examining and interpreting vast amounts of information generated on social media platforms to extract valuable insights. Key aspects of social media data analysis for digital marketing include:

- **Audience Insights:** Social media platforms are rich sources of demographic and behavioral data. Analyzing user interactions, engagement patterns, and preferences helps marketers gain a deeper understanding of their target audience. This insight is crucial for tailoring content and campaigns to specific demographics, thereby increasing relevance and effectiveness.
- **Trend Identification:** Social media data analysis enables marketers to identify current trends and popular topics within their industry or among their target audience. Staying abreast of trends allows for timely and relevant content creation, fostering increased engagement and brand visibility.
- **Competitor Analysis:** By analyzing the social media activities of competitors, businesses can gain valuable insights into their strategies, audience engagement tactics, and content performance. This competitive intelligence helps in refining one's own digital marketing approach and staying ahead in the market.
- **Sentiment Analysis:** Understanding how users feel about a brand, product, or campaign is essential. Sentiment analysis of social media data helps gauge public opinion, identify potential issues, and assess the overall sentiment surrounding a brand. This information guides marketers in crafting more targeted and resonant messages.
- **Campaign Performance Measurement:** Social media analytics allows for the evaluation of campaign performance in real-time. Marketers can track metrics such as reach, engagement, click-through rates, and conversions. This data-driven approach enables quick adjustments to campaigns for optimal results.
- **Customer Feedback and Interaction:** Social media platforms provide direct channels for customers to express their opinions and engage with brands. Analyzing customer feedback and interactions helps businesses address concerns, enhance customer satisfaction, and build a positive brand image.
- **Optimizing Ad Campaigns:** For businesses running paid advertising on social media, data analysis is instrumental in optimizing ad campaigns. It involves refining targeting parameters, ad creatives, and budget allocation based on the performance metrics obtained from social media platforms.
- **Predictive Analytics:** Advanced analytics techniques, such as predictive modeling, can forecast future trends and user behavior. By leveraging historical

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 09, Issue: 06 (November-December, 2023)

data, marketers can make informed predictions about the success of future campaigns and adjust strategies accordingly.

Social media data analysis empowers digital marketers with actionable insights, fostering data-driven decision-making. Harnessing the power of this analysis enhances the overall effectiveness and efficiency of digital marketing efforts in an ever-evolving online landscape.

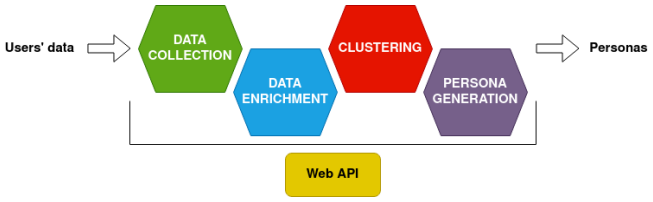

Figure 1: High-level view of the main components

2 Related Work

Predominantly led by Dr. Jim Jansen's team at the Qatar Computing Research Institute, research in this field focuses on the development of *APG (Automatic Persona Generation)*[1], a service that concentrates on automatically generating personas for a YouTube channel using YouTube analytics as the primary source, including data from comments, likes, and view counts categorized by content type and demographics [1]. Other online tools offering similar services emphasize online and social media analytics[2]. Consequently, users of such services are required to have an established online presence (e.g., website or social media accounts) and an analytics tool for monitoring related data (e.g., Twitter Analytics, Facebook Analytics, Google Analytics).

The use of more conventional social networks (beyond YouTube, which specializes predominantly in video content) has been explored to a lesser extent, with limited documentation available. An et al., during the development of personas for Al Jazeera[3], reported the use of Facebook and Twitter data but with certain limitations [2]. Facebook data was restricted to URLs shared by users who followed or discussed Al Jazeera's Facebook page, as the Facebook API requires explicit user consent to access any personal data. The use of Twitter data was also limited to users' biographies, aiming to extract non-behavioral aspects such as occupation and hobbies.

Another valuable source of information is the content within social media posts. Analyzing the text or multimedia content of a user's posts enables one not only to comprehend the author's interests, values, and opinions on a specific topic or product but also to discern their preferred language, the times they are most active, and their tone of voice—factors that are all significant for defining a persona. Various Natural Language Processing (NLP) techniques are available to extract such insights from text. Semantic analysis, for instance, can identify entities such as nouns and corresponding adjectives [3]. Topic analysis can either extract or assign topics, revealing users' topical interests [4]. Sentiment scoring aids in understanding a user's attitude towards a specific topic [4–6]. Deep learning solutions can also be employed to achieve similar results for images and videos [7].

A more popular and well-documented approach is the use of PCA (Principal Component Analysis) [8–10]. PCA is a dimension-reducing algorithm employed to extract information by eliminating non-essential elements with relatively fewer variations. While it was initially used independently [9, 10], recent research leveraging PCA often combines it with clustering algorithms [8].

3 Proposed Approach

The system comprises five primary components, as depicted in Figure 1. These include four logical components situated atop a web API. Each logical component plays a distinct role in executing the necessary steps to transition from raw data—comprising a list of user identifiers sourced from various social networks or other websites—to the ultimate outcome: personas. The sequential steps in this process are as follows:

1. Gather user data from social networks and other websites, including demographic details and activities.
2. Enhance the data by incorporating new insights derived from the outcomes of the preceding step.
3. Group users into similar clusters.
4. Create a persona for each identified cluster.

The initial two steps are interconnected, as data is transmitted to the enrichment module upon collection. The clustering step relies on the enriched data to yield meaningful results, and the persona generator must await the creation of clusters to perform its function. Specifics regarding the design of each component are elaborated in the subsequent sections.

3.1 Data collection

This component is responsible for retrieving user data from the web. Given that the system aims to support the creation of personas across various contexts, the more exten-

[1] persona.qcri.org
[2] www.delve.ai, www.mnemonic.ai
[3] www.aljazeera.com

sive the data collection, the more comprehensive and enriched the final personas will be.

The primary data source is social media data, encompassing conventional platforms such as Twitter, Facebook, or Instagram, as discussed in Chapter **??**. However, the system is designed to accommodate data from non-traditional social media sources (e.g., Strava for sports data, Quora, and Medium for user-specific blogging data). To facilitate this, a comprehensive list of relevant data that can be collected must be defined for each data source. Pertinent user information may include name, location, language, number of followers, profile picture, and activities (user-posted content with metadata like the number of likes, comments, and shares). It is noteworthy that different social media platforms may provide varying sets of data, potentially resulting in data gaps if only one source is used to describe a user. To address this, the system allows the association of multiple data sources with a single user.

A crucial consideration is the periodic collection of user data due to the substantial volume of activities posted on social networks daily (e.g., around 6000 tweets per second on Twitter[4]). Consequently, the collection component should run at regular intervals (e.g., every 24 hours) to maintain up-to-date user data and capture the latest activities.

In general, this component requires a set of user IDs as input, which uniquely identify the user across all associated data sources. The anticipated output includes, for each data source, the corresponding user information and a specified number (n) of user activities. The parameter n, indicating the number of activities to download, can be adjustable but should be judiciously set to prevent system overload. For instance, Twitter allows fetching up to 200 tweets per request.

3.2 Data enrichment

This component is responsible for extracting additional insights from the previously collected data. These insights serve as the basis for clustering users, thereby influencing the characteristics of the final personas. The enrichment component comprises two modules: one focused on enhancing user activities, and the other on enriching user profile information.

3.2.1 Activity enrichment

The objective of this module is to extract valuable insights from activity data, including raw texts and images, along with metadata. Although some insights, such as language or device usage, may be readily available in the collected data, the primary challenge lies in extracting the content's essence. Determining what an activity is about is crucial for understanding user behavior, including their interests and personality. This becomes particularly intricate when dealing with various media types (text, images) and languages.

The anticipated *input* for this module is a user activity, obtained in the previous step. The *output* consists of the enriched activity.

3.2.2 User profile enrichment

The objective of this module is to leverage available data from the collected profile information and enriched activities to extract insights relevant for the final personas. To ensure the system's general applicability, a list of attributes commonly found in online persona templates has been defined. These attributes, detailed in the table below, may not always be extractable, depending on a user's level of activity on a specific social media site and the type of information they share.

The machine learning classifiers can be employed to predict attributes such as gender, age, or type. For instance, gender prediction can be based on a name or a profile picture using computer vision algorithms. Life event detectors may predict a user's marital status and whether they have children by analyzing posts related to marriage or the birth of a child. Personality can be inferred from how the user interacts with others (as discussed in section **??**), and interests result from the entities and topics found in the user's activities.

The expected *input* for this module is the result of collecting user profile information from a data source along with respective enriched activities. The *output* is the enriched user profile.

3.3 Clustering

This component is responsible for creating clusters of similar users based on the characteristics extracted in the previous steps. The *input* is a list of users with enriched attributes, and the expected *output* is structured as follows:

- a mapping of each user to the corresponding cluster index;
- a list of *representative users*, one for each cluster. A representative user is comprised of the characteristics that best define a cluster. It's important to note that it can be either a real user or not, depending on the clustering algorithm. For instance, in the K-Means algorithm, a representative user would be a *centroid*, defined as the average of all the users in a given cluster.

The number of clusters that are found could either be specified as a parameter by the user, or could be automatically determined by the system in order to optimize the quality of the clusters.

[4] `https://www.dsayce.com/social-media/tweets-day/`

3.4 Persona generation

This component is responsible for generating a persona for each cluster identified in the previous step. Given that the output of the clustering component includes a list of representative users, the process involves:

1. Assigning realistic attributes to the representative users, especially if they are not directly linked to real users (e.g., when a representative user is determined as the average of all users in a specific cluster).

2. Provide an identity for each representative user, involving the assignment of a name, photo, and textual description that align with their demographic characteristics.

The *input* is a list of representative users obtained from the clustering stage.The *output* is a list of personas, each tailored to a specific cluster or representative user. Users should ideally have the flexibility to specify which attributes to include in the personas according to their specific requirements.

4 Result Analysis

The results are presented in Table 1. As anticipated, the time required to enrich a single activity remains relatively constant. This duration is calculated as the time difference between receiving the activity in the enrichment queue and saving the enriched activity in the database. As expected, this average is not influenced by the number of activities collected per user, given that each enrichment process is independent of the others. It is solely affected by the external API call and the time taken to store the enrichments in the database.

The total time to enrich a specified number of activities for each of the 90 users is not merely the product of the total activities and the time it takes to enrich a single activity. This deviation is due to the quality of service setting of the pub/sub queue, where setting it to one necessitates additional time as messages must be acknowledged before processing. We refrained from setting it to zero to avoid potential data loss in case of disconnection by some queue clients. It's worth noting that to collect 100 activities, the system requires an entire day. This is due to the activity enrichment API, which permits a maximum of 4500 activities to be fully enriched per day.

Figure 2 illustrates the silhouette scores for various numbers of clusters and three different values of activities per user. The trend indicates that increasing the number of activities per user from 20 to 50 helps refine a user's main interests, consequently reducing the optimal number of clusters from ten to four. The shift from 50 to 100 activities doesn't result in significant changes, leading us to choose 50 as the optimal number of activities per user. This choice also avoids the need to wait for an entire day due to API rate limitations.

Table 1: Performance results for activity enrichment

Operation	Elapsed time
Single activity	0.52 seconds (average)
20 activities per 90 users	16 minutes
50 activities per 90 users	36 minutes
100 activities per 90 users	75 minutes + 24h wait

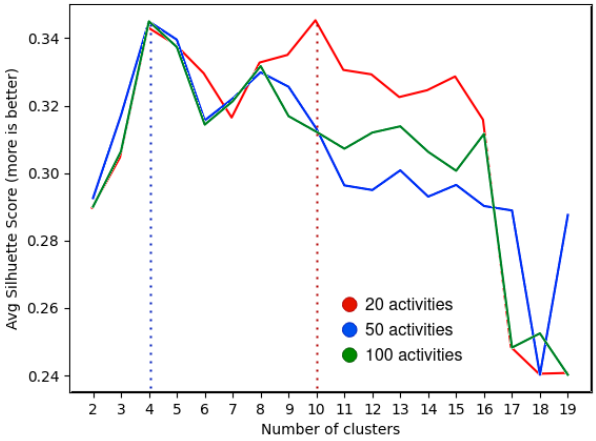

Figure 2: Silhuette score for each number of clusters, for different values of activities per user

This phenomenon occurs because analyzing only a small number of activities increases the risk of misclassifying a user's interests. For instance, during the period when our dataset was compiled, many Italian users tweeted about the UEFA European Football Championship, including politicians and singers. One strategy to prevent classifying users based on outlier activities is to increase the number of activities per user, thus diminishing the impact of outliers. A more sophisticated approach involves tracking the topics users discuss over time, perhaps on a monthly basis. This way, interests that are only prevalent for a limited period could be identified as outliers and excluded from the user's overall interests.

5 Conclusion and Future Work

This research aims to address the research question regarding the viability of automating persona generation from social media data. It proposes a scalable and extensible solution applicable to multiple social media platforms, presenting a prototype specifically tailored to Twitter data.

Distinguishing itself from existing work, this thesis contributes to the current state of the art by thoroughly exploring all phases of the process rather than focusing solely on clustering. The enrichment component incorporates innovative techniques, such as leveraging Wikipedia content to comprehend user discourse. This approach supports a wide array of languages and utilizes all components of an activity (text, images, external links, etc.) for classification purposes. The proposed system architecture and data models are designed for general-purpose personas, yet they can accommodate the creation of sector-specific personas by introducing new classifiers. Moreover, the stream processing design minimizes bottlenecks and achieves a high level of separation of concerns.

References

[1] J. An, H. Kwak, S. Jung, J. Salminen, M. Admad, and B. Jansen, "Imaginary people representing real numbers: Generating personas from online social media data," *ACM Transactions on the Web (TWEB)*, vol. 12, no. 4, pp. 1–26, 2018.

[2] J. An, H. Cho, H. Kwak, M. Z. Hassen, and B. J. Jansen, "Towards automatic persona generation using social media," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*. IEEE, 2016, pp. 206–211.

[3] D. H. Maulud, S. R. Zeebaree, K. Jacksi, M. A. M. Sadeeq, and K. H. Sharif, "State of art for semantic analysis of natural language processing," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 21–28, 2021.

[4] S. Huang, W. Peng, J. Li, and D. Lee, "Sentiment and topic analysis on social media: a multi-task multi-label classification approach," in *Proceedings of the 5th annual ACM web science conference*, 2013, pp. 172–181.

[5] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment analysis on social media," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2012, pp. 919–926.

[6] A. Mewada, R. K. Dewang, P. Goldar, and S. K. Maurya, "Sentibert: A novel approach for fake review detection incorporating sentiment features with contextual features," in *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, ser. IC3-2023. Association for Computing Machinery, 2023, pp. 230–235. [doi: https://doi.org/10.1145/3607947.3607991

[7] P. Rodríguez, D. Velazquez, G. Cucurull, J. M. Gonfaus, F. X. Roca, S. Ozawa, and J. Gonzàlez, "Personality trait analysis in social networks based on weakly supervised learning of shared images," *Applied Sciences*, vol. 10, no. 22, p. 8170, 2020.

[8] J. Salminen, K. Guan, S.-G. Jung, and B. J. Jansen, "A survey of 15 years of data-driven persona development," *International Journal of Human–Computer Interaction*, pp. 1–24, 2021.

[9] R. Sinha, "Persona development for information-rich domains," in *CHI'03 extended abstracts on Human factors in computing systems*, 2003, pp. 830–831.

[10] N. Tu, Q. He, T. Zhang, H. Zhang, Y. Li, H. Xu, and Y. Xiang, "Combine qualitative and quantitative methods to create persona," in *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering*, vol. 3. IEEE, 2010, pp. 597–603.