Chapter 8
Persona Classification of Celebrity Twitter Users

Aastha Kaul, Vatsala Mittal, Monica Chaudhary, and Anuja Arora

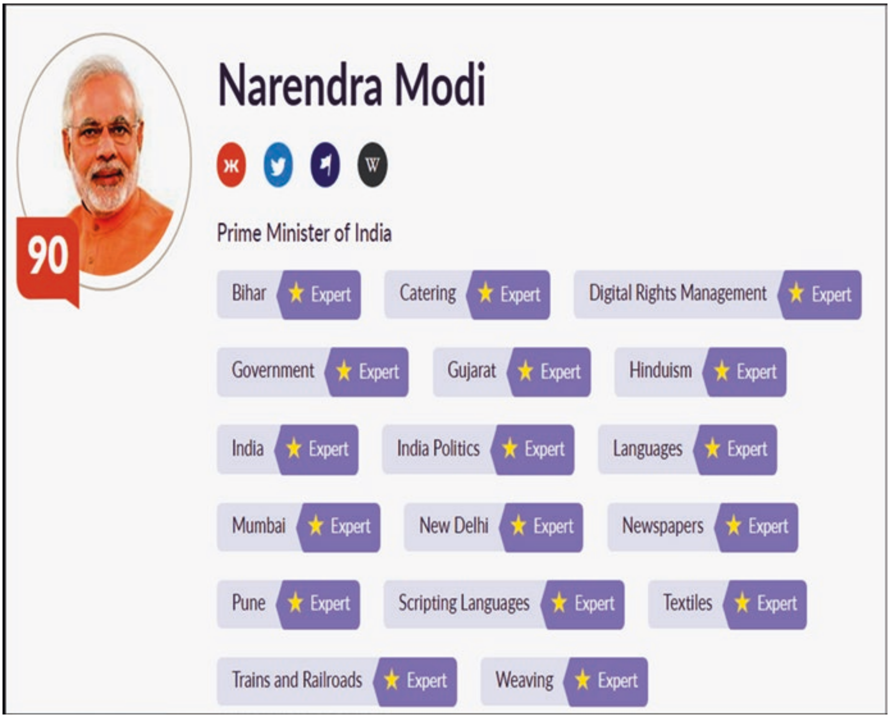
8.1 Introduction

Social networking on the web has dramatically grown over the last decade. Twitter is one of the most popular online social networking sites where many celebrities post tweets for their fans and also post something related to an event. Twitter is a microblogging service because it enables users to send and read a short text message, which is known as "tweet." There are 316 million monthly active users on Twitter, and 500 million tweets are posted per day (internetlivestats.com). Through self-description, status update, and tweets, we can find a lot about the users. We can use these tweets to analyze the interest of users and get to know the trends going on at any place. A user's knowledge of social sites could be remarkably improved if other information like demographic attributes and user's personal interest and the interest of other users are considered. This is truer in case of celebrity users. This chapter attempts to analyze celebrity tweets to provide relevant recommendations to the practitioners. Such analysis may help in designing a smart recommendation system. Few websites contain similar sort of features and show users' interest areas such as klout, which gives klout score (out of 100) to every twitter user according to the number of twitter posts and the post's influential content shared by them. It also shows all the influential topics of a specific twitter user, as shown in Fig. 8.1 for the Indian Prime minister; Mr. Narendra Modi's klout score is "90." .

With the rapid increase of information on these social media sites, it is becoming difficult for users to get the most relevant tweets and for companies to target the most relevant users according to a particular topic/theme. Therefore, system requires research methods to extract influential topics and to validate accuracy of influential topic identification, which itself is a challenging task. Influential topic identification can help various applications and covers various perspectives:

A. Kaul · V. Mittal · M. Chaudhary · A. Arora (✉)
Jaypee Institute of Information Technology, Noida, India

N. P. Rana et al. (eds.), *Digital and Social Media Marketing*, Advances in Theory and Practice of Emerging Markets, https://doi.org/10.1007/978-3-030-24374-6_8



Narendra Modi

90

Prime Minister of India

Fig. 8.1 Klout score and influential topic identification

- *Users' perspective*: user can get relevant tweets according to their choice of influential topic. For example: Gujurat is shown as one of the influential topic in Fig. 8.1. Even, User can further explore influential subtopics under the influential broad topic.
- *Company perspective*: while going for brand promotion companies may target users according to a specific influential topic instead of users who are not at all interested in product. Hence, this may help companies to target specific influential topic users for their brand promotion as well.
- *Twitter perspective*: Twitter itself can use influential topic as feature to design user's timeline/news feed. Tweets order can be decided based on users' influential topic for users news feed.

This research work is an effort in this direction, and it tries to identify the persona of a user based on their twitter feeds. The user's tweet feed-based influential topic has been identified using Latent Dirichlet Allocation (LDA) algorithm and the results have been refined using hypernyms. Three classification algorithms (Naïve Bayes, Decision Tree, and Support Vector Machine) are used to classify the users' persona, out of which Naïve Bayes provides the highest accuracy. In context to literature study, this work makes novel and unique research contributions. The refined research objectives are:

1. Critically analyze the existing classification methods for persona classification, the techniques and key theoretical contributions.
2. Design an approach to classify user's persona through tweets' contents into six predefined categories.
3. Identify persona of selected celebrities and users according to posted tweets and retweets content.

This chapter attempts to analyze celebrity tweets to provide relevant recommendations to the practitioners. The tweets of celebrity users are classified using two distinct approaches (1) Fixed Classification into six predefined categories and (2) Generating a category if the tweet does not belong to any defined category. The first kind of classification has been done in three different ways; by individually applying Naïve Bayes, Decision Tree and Support Vector Machine. For generating a new category Latent Dirichlet Allocation is used. Henceforth, this Persona Classification of Celebrity Twitter Users will help users to gain insight into their interests thereby decluttering their twitter feed and showing them relevant content on their feed. With an understanding of celebrity persona, smart recommendation systems can also be designed. The chapter is organized as follows: Integrated summary of related literature is detailed in Sect. 8.2, which further divides literature in two subsections—tweets classification and celebrity persona. Dataset statistics are enlisted in Sect. 8.3 which covers data preprocessing also. Section 8.4 discusses user persona research method. Experimental evaluation and results are summarized in Sect. 8.5, where results are presented of proposed approach in order to reflect performance of the developed system. Finally, Sect. 8.6 deals with the implications and limitations of the study, followed by the conclusion section.

8.2 Related Literature

Twitter is a popular social networking site, where users search for social information such as breaking news, posts about celebrities, and trending topics. Since Twitter's launch, its popularity has been increasing. It has been used in various campaigns, elections and as a news medium, and therefore it is important to classify tweets into general categories for better information retrieval and easier understanding of topics.

Internet users share private content, such as personal information or photographs, even their likes and dislikes. An individual's online behavior creates their unique persona and this persona helps businesses to analyze their behavior patterns and needs and deliver accordingly. By classifying a user's tweet into general categories like sports, politics, and entertainment, we define the user's "persona". If a user mostly tweets about sports, we can say he is interested in sports and thereby define his or her persona accordingly. Several works have been done in the field of social networking (Shiau et al. 2017), namely, classification of gender (Ugheoke and Saskatchewan 2014), classification of the topic (Sriram et al. 2010), sentiment analysis of Twitter users based on tweets (Go et al. 2009), event detection (Sakaki et al. 2010), and community detection, which provide us an insight into the user's interests and generate their personality.

8.2.1 Twitter Classification

Twitter Classification approach has been evolving since the year 2000. This is very much needed as twitter is not just a social networking site but rather a powerful medium to express your thoughts and opinions; through these tweets we can find a lot about users.

In 2000 text categorization was done using basic machine learning algorithm like Support Vector Machine (SVM) (Siolas and d'Alché-Buc 2000). The authors proposed to solve a text categorization task using a new metric between documents, based on a priori semantic knowledge about words. This metric can be incorporated into the definition of radial basis kernels of Support Vector Machines or directly used in a K-nearest neighbor algorithm. The method proposed was based on the exploitation of the information provided by Wordnet. They found out that semantic smoothing is relevant for text categorization and the introduction of the semantic proximity matrix in the kernel increases the number of support vectors. Moreover, in case of SVM, the results in terms of precision, recall, and accuracy appear to be very high. Using Wordnet there was another study by Elberrichi in 2008 (Elberrichi et al. 2008). The approach in this study was composed of two stages. The first stage relates to the learning phase which was to merge terms with associated concepts to represent texts. The second stage relates to the classification phase which consisted of generating the weighted vector for all categories and then using a similarity to find the closest category. They reached an f1 score of 71.7% which in comparison of Bag of Words representation was an increase by at least 6% on the Reuters Data.

Short text classification in twitter was done in 2010 to improve information filtering by proposing a method to classify the text into a predefined set of generic classes such as news events etc. (Sriram et al. 2010). In this approach, the learning model trains itself using these features. They proposed that categorization of tweets need prior knowledge of the tweet such as corporate tweets have different motivation as compared to that of a personal tweet. The results showed that the author feature was an improvement over simple Bag of Words classification. When all the features were applied there was an overall of 50% improvement compared to the simple bag of words classification. The results also show that noisy data may degrade the performance of the proposed approaches and hence noise removal is important.

Another very interesting study was conducted in 2011 by Golbeck, where the big five personality model was administered into tweets using regression (Golbeck et al. 2011). For prediction personality of a user, this new approach proposed to bridge the gap between social media and personality research by using the information people reveal in their online profiles. They administered the Big Five Personality Inventory to 279 subjects through a Twitter application. This model contained five personalities as follows: (1) Openness to Experience (2), Conscientiousness (3), Extroversion (4), Agreeableness (5), and Neuroticism. To analyze the data, they used mainly two tools: first was LIWC (Linguistic Inquiry and Word Count). Second was that they then ran the text again on the MRC Psycholinguistic Database. Next task was to run Pearson Correlation analysis between the user's personality score and each of the feature obtained from analyzing the tweets. Finally, predicting personality was done by using regression tools on WEKA.

In 2012 Lehmann explained Dynamical classes of collective attention in Twitter by analyzing tweets and finding the evolution of hash tag popularity over time defining discrete classes of hash tags (Lehmann et al. 2012). They focused their analysis on those hashtags that exhibited a popularity peak during our observation period and systematically analyze the corresponding messages ("tweets") by grounding the words. On visual inspection the individual temporal profiles of hashtag usage display behaviors that typically fall into one of the following three categories: continuous activity, periodic activity, or activity concentrated around an isolated peak. For identifying the activity peaks for every hashtag they computed time series of daily activity. To correlate the temporal activity patterns with the content, they performed semantic grounding of tweets using Wordnet. For identifying classes they use a standard implementation of the Expectation Maximization (EM) algorithm (Fraley and Raftery 2006). They then used Wordnet to systematically analyze contents of tweets associated with the group of hashtags. LDA was also used for short text classification by Chen in 2016 in which he gave an improved short text classification method based on Latent Dirichlet Allocation topic model and K-nearest neighbor algorithm is proposed (Chen et al. 2016). In addition, it presents a topic similarity measure method with the topic-word matrix and the relationship of the discriminative terms between two short texts. LDA topic model is employed to generate the topic word matrix. Feature vectors are extracted for a sample of short texts; reweight the feature vector using the topic-word matrix through discriminative words and then, calculate the topic similarity between two short texts. At last, classifiers are trained by the labeled dataset. The topic similarity combines the semantic features generated by LDA topic model and the information of discriminative words. Therefore, they exploit the topic similarity as distance metric of KNN algorithm. The Precision Recall and F1-measure of their method have significantly increased by 25–47% over KNN. Figure 8.2 shows the summarized past work done on Tweet Classification by various researchers in the field of Twitter feeds as social media content.

8.2.2 Using Celebrity Persona

According to Rein, Kottler, and Stoller, celebrity refers to an individual "whose name has attention-getting, interest-riveting and profit generating value" (Kotler et al. 1987). Celebrities have always served as beacons of the mass public. Celebrities

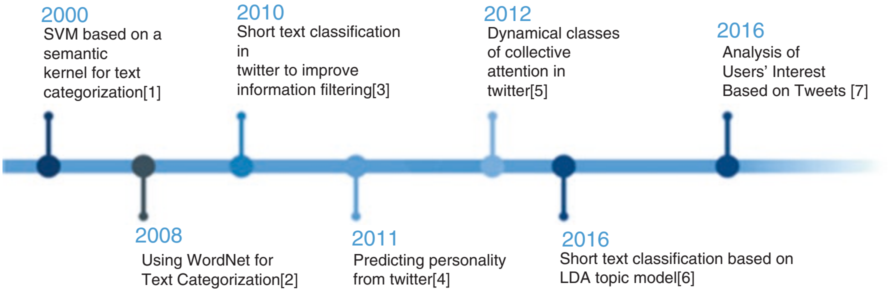

2000
SVM based on a semantic kernel for text categorization[1]
2010
Short text classification in twitter to improve information filtering[3]
2012
Dynamical classes of collective attention in twitter[5]
2016
Analysis of Users' Interest Based on Tweets [7]
2008
Using WordNet for Text Categorization[2]
2011
Predicting personality from twitter[4]
2016
Short text classification based on LDA topic model[6]

Fig. 8.2 Tweet classification literature summary (Research papers with publishing years)

with their unique persona help defining the spirit of any particular moment that relied in part on its intervention through film, radio, popular music and television (Marshall 2010). Celebrity taught generations how to engage and use consumer culture to "make" oneself (Leiss et al. 1990). With the advent of social media celebrities' public self is presented through a new layer of interpersonal conversations. Celebrity use of social media articulates this change (Marshall 2010). Celebrity practitioners reveal what appears to be personal information to create a sense of intimacy between participant and follower, publicly acknowledge fans, and use language and cultural references to create affiliations with followers (Marwick and Boyd 2011).

Celebrity persona is a site of tension and ambiguity in which an active audience has the space to make meaning of their world by accepting or rejecting the social values embodied by a celebrity image (Balasubramanian et. al. 2016). Thus, an examination of celebrity persona and the social meaning and significance generated by their persona offers new ways of understanding the society and markets (AlAlwan et al. 2017; Kapoor et al. 2018). The celebrity persona is not confined to their professional image but actually consists of everything publicly available about them (Dyer 2013).

Armstrong was probably the first one to analyze "celebrity persona" as a property (Armstrong Jr 1990). In his study, he very creatively explained how celebrity persona has become property, how the gradual accretion of characteristics such as the right to exclude and alienate usually associated with property. There are many quantitative studies and case studies that explore the use of celebrities in propagating business and brands. One such case study was about Jamie Oliver (well known as television celebrity the Naked Chef) into the promotions of one of Britain's leading grocery chains, which involves a high profile campaign that has been adopted in order to imbue the company's products with an image of quality (Byrne et al. 2003). Another study by Meyers in 2009 explores the power of popular media in shaping a celebrity in the case of the famous singer Britney Spears (Meyers 2009). There are some studies that also analyze celebrity persona as an important means of delivering politics via the mass media. This particular study analyzes Arnold Schwarzenegger's persona in his speeches (Drake and Higgins 2006). In another interesting study by (Marshall 2010), social media via social network is seen as a “presentational media’ for celebrities. Social media is also a form of presentation of the self and produces this new hybrid among the personal, interpersonal, and the mediated. Via Facebook, MySpace, Friendster, and Twitter individuals engage in an expression of the self, which is like the celebrity discourse of the self (Marshall 2010).

Celebrity endorsement in business is a popular advertising technique (Dwivedi et al. 2015; Shareef et al. 2019). Celebrity endorsement advertising has been recognized as “a ubiquitous feature of modern marketing” (McCracken 1989). There have been few studies examining the celebrity advertising in their respective country; USA (Stephens and Rice 1998), Japan (Kilburn 1998), United Kingdom (Davies and Slater 2015), China (Jiang et al. 2015), India (Agnihotri and Bhattacharya 2016), Kenya (Njuguna and Otieno 2015), Mexico (Felix and Borges 2014), and Australia (Dixon et al. 2014).

8.3 Research Methodology

Perhaps the best way to illustrate the theoretical claims is to examine how they apply to a specific celebrity persona. That is what this chapter is offering. The chapter attempts to examine the tweets by six celebrities and hence analyzes their persona.

With internet availability and reach, massive amount of celebrity-focused media is available. It would be tough to select a “celebrity” as there is very large number of people who qualify as “celebrities.” But safely, it can be said that not all celebrities are equal in terms of media coverage. One defining characteristic of celebrity is that a social actor attracts large-scale public attention: the greater the number of people who know of and pay attention to the actor, the greater the extent and value of that (Rindova et al. 2006). Another important characteristic of a celebrity is that the actor elicits positive emotional responses from the public (Heider 1946; Trope and Liberman 2000).

So, for this study, it is necessary to focus on a celebrity whose image is easily traceable and is active (textual) on twitter. The study here examines the celebrity persona of six celebrities; Ellen Degeneres, Bill Gates, Barack Obama, Dalai Lama, Amitabh Bachchan, and Selena Gomez. The tweets of these celebrities are captured during the year 2017. These tweets are then classified into six distinct categories; Education, Entertainment, Health, Nature, Politics, and Sports.

Broad research workflow is presented in Fig. 8.2. As we see from the figure, complete research work is classified in two parts:

1. User Persona Classification Method: In this method, three classification algorithms have been implemented to categorize celebrities’ tweets into celebrity persona.

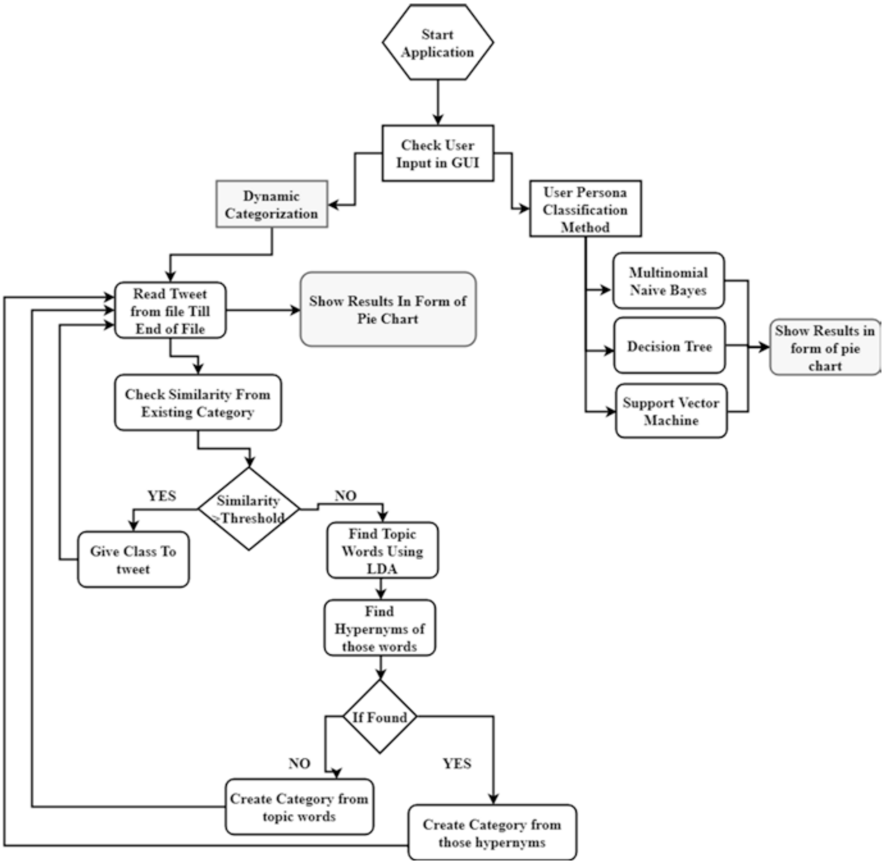

Start Application
Check User Input in GUI
Dynamic Categorization
User Persona Classification Method
Multinomial Naive Bayes
Decision Tree
Support Vector Machine
Show Results in form of pie chart
Read Tweet from file Till End of File
Show Results In Form of Pie Chart
Check Similarity From Existing Category
YES
Similarity >Threshold
NO
Give Class To tweet
Find Topic Words Using LDA
Find Hypernyms of those words
If Found
NO
YES
Create Category from topic words
Create Category from those hypernyms

Fig. 8.3 Research workflow

2. Dynamic Categorization: In this method, a new category of persona based on tweet content has been generated by the system using LDA topic modeling algorithm and hypernyms of words (Fig. 8.3).

8.3.1 Data Statistics

Different categories' twitter feed data characteristics that were used during the work are detailed in Table 8.1. This helps the reader to identify the present variation in fetched data and to validate the usefulness, correctness of applied approaches towards varying Twitter data. Data has been collected using Twitter API (Tweepy for Python). Twitter feed data for different celebrity users with their characteristics is shown in Table 8.2.

Table 8.1 Dataset statistical information for Twitter Feed Data for six defined categories

Category	#of tweets	Average no. of tweets	Max. length	Avg. length	Min. length
Education	2370	0.088	399	128	19
Entertainment	4883	0.183	370	127	10
Health	2500	0.094	283	126	17
Nature	1933	0.072	126	29	0
Politics	7500	0.281	376	127	9
Sports	7500	0.281	357	125	6

Table 8.2 Dataset information about User’s Twitter Feed Data for six different users

Name of user	#of tweets	Average no. of tweets	Max. length	Avg. length	Min. length
Ellen Degeneres	3210	0.193	133	94	34
Bill Gates	2489	0.150	153	125	18
Barack Obama	3215	0.194	157	116	28
Dalai Lama	1286	0.775	150	115	30
Amitabh Bachchan	3160	0.190	324	85	4
Selena Gomez	3224	0.194	196	84	2

8.3.2 Data Preprocessing

Tweets are a new genre of text, which are short, informal, ungrammatical, and noise prone. So, to add a bit of a structure and to make the text more readable and cleaner, various steps are followed which are as follows:

Tokenization and stop words removal: Stop words do not give vital information in the understanding of a text. Hence they can be removed in order to perform a better analysis of data. For example we have a tweet “Bruno Mars New Album To Drop On 2016? I’m On A Mission”, what tokenization and removal of stop words will do is that words like “on”, “a”, “to” and “I’m” will be removed and the resultant tweet will be “Bruno Mars New Album Mission”.

Emoticons removal: When a tweet is extracted using Twitter API and stored in .csv format, the emoticons don’t appear as they are posted. Instead, they get converted and require to be encoded using utf-16and hence should be removed. For example “Weekend is looking Great ðŸ‡©ðŸ‡ªðŸ GO OFF ðŸ ðŸ ðŸ ðŸ ðŸ Via: @mikealdred #e30 #off #rally #xi #winter #gooff #euro #bmwworld #performance”. It can be clearly seen the symbols are a waste and therefore should be removed.

Punctuation marks removal: In tweet classification task punctuation marks don’t prove to be useful and therefore can be removed. For example in the tweet “Your Favourite Singers Will Be Back For American Idol’s Final Season!!!!!” The exclamation marks don’t provide any additional information and hence should be removed.

Hashtag and hyperlink removal: The hashtag symbol # and hyperlinks are treated as waste as we don't extract any information from them and hence their removal helps in getting better results. For example "Watch: Elton John on 'Ellen' talks sons and new album https://t.co/V8TcEe1mCI #DeGeneres" is transformed into "Watch Elton John on 'Ellen' talks sons and new album DeGeneres."

Stemming and lemmatization: Stemming and lemmatization were used to reduce inflectional forms and derivationally related forms of a word to a common base form.

POS tagging: While performing POS tagging, it is observed that the various tags are activated over the lemmatized words such as -Noun (*N*) tag, Adjective (*ADJ*) tag, Verb (*VB*) tag, Adverb (*ADV*) tag, and Unknown (*UNK*) tag. *UNK* tag refers to that tag for which no POS category is provided to the tokens, since such tokens are not listed in POS tagging list. Tagging is done to extract only proper nouns and nouns from the tweets.

8.3.2.1 User Persona Classification Method

In this study, classification has been done in two phases. First phase is category-specific persona classification, which helps to assign persona of a user according to their posted tweets. This persona category is already defined because data from six predefined categories have been extracted. Even, this phase is helpful to validate results as well due to existing persona category. System is able to justify the accuracy in detecting persona based on terms usage in tweets. Second phase gives persona, which is out of these six predefined categories. Latent Dirichlet Allocation and hypernyms has been used to generate persona category.

Category-specific user persona classification: This is the first kind of tweet classification and has been done in three different ways; by individually applying Naïve Bayes, Decision Tree and Support Vector Machine. Six predefined categories namely Education, Entertainment, Health, Sports, Nature and Politics were used. All the models were first trained with 80% tweets per category and then tested with the rest. The techniques were then evaluated and accuracy for each classifier was calculated. Celebrity user's Twitter data was then used as input for different trained models to find out the resultant categories of their tweets and the results were shown in form of a pie chart.

Proceeding with one tweet at a time from, similarity with each category with the help of word similarity is calculated. If this similarity was above the defined threshold of 0.21 then the category with maximum similarity to the tweet was assigned. But if the similarity was less than the threshold, topic modeling using LDA was done on the tweet. Hypernyms (Root Word) of the topic words generated using LDA were found and appropriate hypernyms were made the category of the tweet.

Dynamic categorization: For the second kind of classification, we had text files containing a minimum of 2000 different words for all the six categories. Taking one tweet at a time from the celebrity user's twitter data, similarity with these words were calculated using TF-IDF. If this similarity was above the defined threshold, then the category with the maximum similarity was assigned. But if the similarity calculated was below the defined threshold a new category for the tweet had to be generated. For this purpose, Latent Dirichlet Allocation was used to generate topic words for each tweet and using Wordnet library hypernyms (Root Word) were found for the topic words. These hypernyms were then made as the category of the tweet and a text file was created in which all the related words were added, if it didn't exist already. Now when the next tweet was considered similarity was calculated with the words of the six defined categories and the words of the newly generated classes. For proper nouns as no hypernyms were generated, the proper noun itself had to be made the new category of the tweet. The results for the celebrity users in this kind of classification were again shown in form of a pie chart.

8.4 Experimental Evaluation and Results

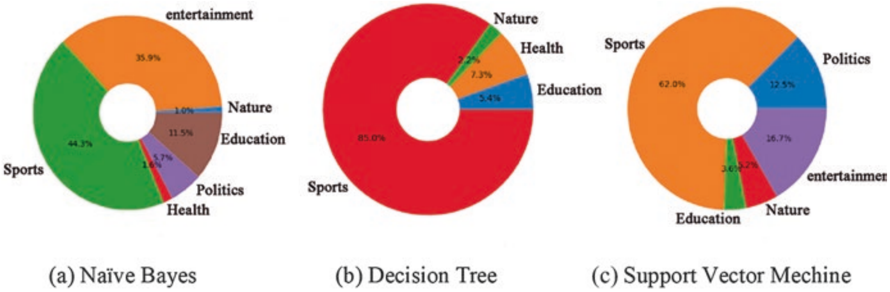
8.4.1 Category-Specific User Persona Classification Results

Table 8.3 shows a comparison between used supervised techniques like Naïve Bayes, Decision Tree, and SVM. It can be seen that Naïve Bayes performs the best with an accuracy of 90.02%, while Decision Tree performs the worst with an accuracy of 44.69%.

Figure 8.4a shows the classification of tweets of user "Ellen Degeneres" into the defined six categories using Naive Bayes Algorithm. It shows that 44.3% of the tweets belong to Sports category, 35.9% fall into Entertainment while only 1.4% belongs to Health. Figure 8.4b uses Decision Tree for classification and reports that 85% of the tweets fall into the Sports category; 2.2% belong to nature and there are tweets in the categories Entertainment and Politics. Figure 8.4c shows the classification using SVM algorithm showing that 62% of the tweets are under Sports, 16.7% fall into Entertainment while only 2.2% fall under Nature. Similarly analyses of "Barak Obama" are shown in Fig. 8.5.

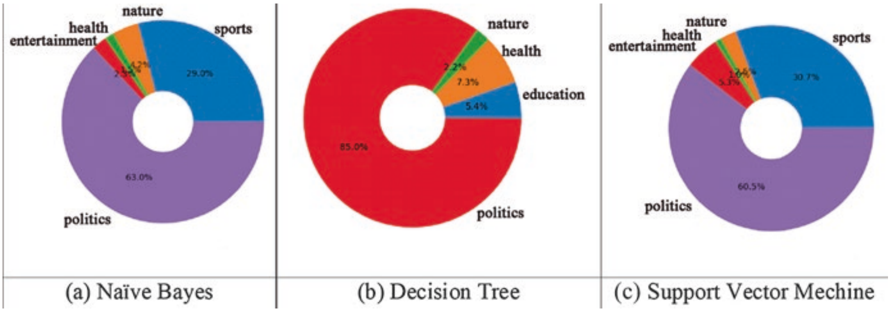
Table 8.3 Comparison of various supervised learning techniques

Classification algorithm	Accuracy
Naive Bayes	90.02%
Decision Tree (Gini Index)	44.69%
Support Vector Machine	61.32%



(a) Naïve Bayes (b) Decision Tree (c) Support Vector Mechine

Fig. 8.4 “Ellen Degeneres” Persona classification results



(a) Naïve Bayes (b) Decision Tree (c) Support Vector Mechine

Fig. 8.5 “Barack Obama” Persona classification results

8.4.2 *Dynamically Identified Persona Results*

Figure 8.6 shows classification of a celebrity tweets which do not fall into the predefined categories. So, a new persona category is automatically generated by the system. The result for “Ellen Degeneres” shows that 38.6% of the tweets fall into Entertainment, 26% belong to Education, while only 2.5% belong to Sports. Other than the predefined categories, newly generated category for Degeneres accounts for 9.7% of her tweets; Instagram having a share of 0.6%; Talk show having 1.3%, and Clinton accounting for 1.6% of the tweets and some more. Similarly analyses of “Barak Obama” are shown in Fig. 8.7.

8.5 Implications and Limitations

To have an insight into a user’s interest and personality can prove to be beneficial in various domains. The important implications of this empirical study are multifold. It may be very useful for academics to go deep into the study of online persona of

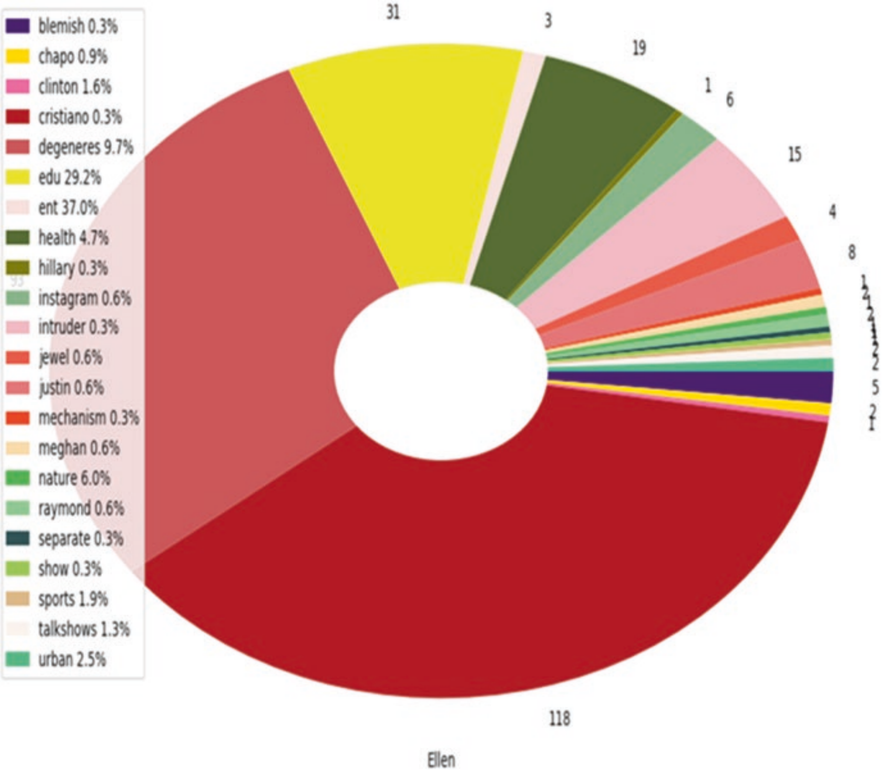


Fig. 8.6 Generated categories and classification for "Ellen Degeneres"

celebrities and how it impacts businesses. The study is also helpful for marketers and practitioners. With the understanding of celebrity persona, smart recommendation systems can be designed. For an online user, the study can be helpful by telling them which celebrity to follow/not to follow; thereby helping them declutter twitter feed. For big brands and companies, this study can be helpful in strategic placement of advertisements according to a user's area of interests. Also, by understanding celebrity persona, business can be benefitted by strategically deciding which celebrity to hire for brand endorsements.

One of the biggest problems encountered was that the data extracted from twitter was informal, abbreviated, and contained lots of symbols; so obtaining useful words and meaning of the sentence was the biggest challenge. Hence various data preprocessing techniques were to be applied in order to give structure to the data. The study is also limited to English language only and thus words of any other language were removed.

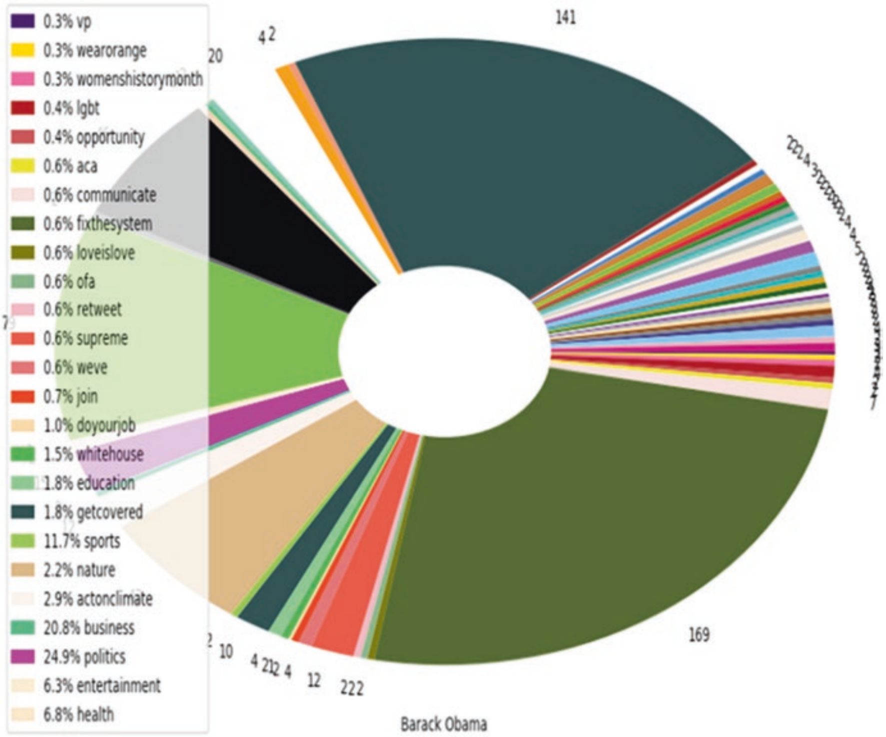


Fig. 8.7 Generated categories and classification for "Barak Obama"

8.6 Conclusion

Twitter is a way to express and share your views and thoughts application, which allows its user to share pictures and videos and post tweets. In this chapter, the objective was to classify a celebrity user's tweet and generate their interests so as to define their persona. In order to do this, two approaches were used. In the first approach, the most basic and common supervised learning techniques were used; Naïve Bayes, Decision Tree, and SVM to classify a user's tweet into six predefined categories: Education, Entertainment, Sports, Nature, Politics, and Health. In the second approach, the tweets were classified not just into six predefined categories but were put in a new category if the tweet did not fit in the defined categories, i.e., it didn't meet the defined threshold. Using Latent Dirichlet allocation, topic words were found in the tweet and then new categories were defined using hypernyms of the obtained topic words. Both the techniques help us get an overview of the user's interests and define their persona.

References

Agnihotri, A., & Bhattacharya, S. (2016). Celebrity endorsement and market valuation: Evidence from India. In *Celebrating America's pastimes: Baseball, hot dogs, apple pie and marketing?* (pp. 709–713). Cham: Springer.

AlAlwan, A., Rana, N. P., Dwivedi, Y. K., & Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics, 34*(7), 1177–1190.

Armstrong, G. M., Jr. (1990). The reification of celebrity: Persona as property. *Louisiana Law Review, 51*, 443.

Balasubramanian, P., Gopal, A. V., & Reefana, S. (2016). A case study on misleading celebrity endorsements and its impact on consumer behavior. *Bonfring International Journal of Industrial Engineering and Management Science, 6*(3), 93–95.

Byrne, A., Whitehead, M., & Breen, S. (2003). The naked truth of celebrity endorsement. *British Food Journal, 105*(4/5), 288–296.

Chen, Q., Yao, L., & Yang, J. (2016). Short text classification based on LDA topic model. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)* (pp. 749–753). Washington: IEEE.

Davies, F., & Slater, S. (2015). Sport celebrity endorsement and the British consumer. In *Marketing dynamism & sustainability: Things change, things stay the same…* (pp. 191–191). Cham: Springer.

Dixon, H., Scully, M., Niven, P., Kelly, B., Chapman, K., & Donovan, R. (2014). Effects of nutrient content claims, sports celebrity endorsements and premium offers on pre-adolescent children's food preferences: Experimental research. *Paediatric Obesity, 9*(2).

Drake, P., & Higgins, M. (2006). I'ma celebrity, get me into politics. In S. Holmes & S. Redmond (Eds.), *Framing celebrity: New directions in celebrity culture* (pp. 87–100). London: Routledge.

Dwivedi, Y. K., Kapoor, K. K., & Chen, H. (2015). Social media marketing and advertising. *The Marketing Review, 15*(3), 289–309.

Dyer, R. (2013). *Heavenly bodies: Film stars and society*. New York: Routledge.

Elberrichi, Z., Rahmoun, A., & Bentaalah, M. A. (2008). Using WordNet for text categorization. *International Arab Journal of Information Technology (IAJIT), 5*(1).

Felix, R., & Borges, A. (2014). Celebrity endorser attractiveness, visual attention, and implications for ad attitudes and brand evaluations: A replication and extension. *Journal of Brand Management, 21*(7-8), 579–593.

Fraley, C., & Raftery, A. E. (2006). *MCLUST version 3: An R package for normal mixture modeling and model-based clustering*. Washington: Washington University Seattle Department of Statistics.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (Vol. 1, issue 12).

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)* (pp. 149–156). Washington: IEEE.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology, 21*(1), 107–112.

Jiang, J., Huang, Y. H., Wu, F., Choy, H. Y., & Lin, D. (2015). At the crossroads of inclusion and distance: Organizational crisis communication during celebrity-endorsement crises in China. *Public Relations Review, 41*(1), 50–63.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers., 20*(3), 531–558.

Kilburn, D. (1998). Star power. *Adweek Eastern Edition, 39*(2), 20–21.

Kotler, P., Rein, I. J., & Stoller, M. R. (1987). *High visibility*. New York: Dodd, Mead & Company.

Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 251–260). New York: ACM.

Leiss, W., Kline, S., & Jhally, S. (1990). *Social communication in advertising: Persons, products & images of well-being*. New York: Psychology Press.

Marshall, P. D. (2010). The promotion and presentation of the self: Celebrity as marker of presentational media. *Celebrity Studies, 1*(1), 35–48.

Marwick, A., & Boyd, D. (2011). To see and be seen: Celebrity practice on Twitter. *Convergence, 17*(2), 139–158.

McCracken, G. (1989). Who is the celebrity endorser? Cultural foundations of the endorsement process. *Journal of consumer research, 16*(3), 310–321.

Meyers, E. (2009). "Can you handle my truth?": Authenticity and the celebrity star image. *The Journal of Popular Culture, 42*(5), 890–907.

Njuguna, S. P., & Otieno, H. N. (2015). Influence of celebrity endorsements on young consumers' brand recall behaviour in Kenya: A case of Nairobi County.

Rindova, V. P., Pollock, T. G., & Hayward, M. L. (2006). Celebrity firms: The social construction of market popularity. *Academy of Management Review, 31*(1), 50–71.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 851–860). New York: ACM.

Shareef, M. A., Mukerji, B., Dwivedi, Y. K., Rana, N. P., & Islam, R. (2019). Social media marketing: Comparative effect of advertisement sources. *Journal of Retailing and Consumer Services, 46*, 58–69.

Shiau, W.-L., Dwivedi, Y. K., & Yang, H.-S. (2017). Co-citation and cluster analyses of extant literature on social networks. *International Journal of Information Management, 37*(5), 390–399.

Siolas, G., & d'Alché-Buc, F. (2000). Support vector machines based on a semantic kernel for text categorization. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on IEEE* (Vol. 5, pp. 205–209). Washington: IEEE.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 841–842). New York: ACM.

Stephens, A., & Rice, A. (1998). Spicing up the message. *Finance Week, 76*(26), 46–47.

Trope, Y., & Liberman, N. (2000). Temporal construal and time-dependent changes in preference. *Journal of Personality and Social Psychology, 79*(6), 876.

Ugheoke, T. O., & Saskatchewan, R. (2014). *Detecting the gender of a tweet sender*. M.Sc. Project report, Department of Computer Science, University of Regina, Regina.

Aastha Kaul is an M.Tech student of Computer Science Engineering at Jaypee Institute of Information Technology. Her current field of research is analysis of different social media platforms. She has published a paper "Multivariate Features Based Instagram Post Analysis to Enrich User Experience" in the ITQM conference held at Jaypee Business School in 2017. She is interested in the field of machine learning and artificial intelligence.

Vatsala Mittal is an M.Tech student studying in the field of Computer Science Engineering at Jaypee Institute of Information Technology. Her current field of research is analysis of textual data from social media websites. She is interested in data sciences with focus on analysis of images, audio files and textual data. She has published a paper "Multivariate Features Based Instagram Post Analysis to Enrich User Experience" in the ITQM conference held at Jaypee Business School in 2017.

Monica Chaudhary is PhD in Management (Marketing) and double masters in Management and Economics with more than 14 years' experience in Academics, Research, Consulting and Industry.

Currently she is working as an Assistant Professor, Humanities, with Jaypee Institute of Information Technology, Noida, India. She is an expert in Marketing and Consumer Behavior. She has authored several refereed research papers in reputed research journals (SSCI & SCOPUS indexed, ABDC listed). She is also Editorial Board Member & Reviewer for leading peer-reviewed research journals and conferences. She has chaired many conference sessions, delivered talks, and conducted workshops.

Dr. Anuja Arora is working as Associate Professor in Computer Science Engineering department of Jaypee Institute of Information Technology. She has an academic experience of 14 years and industry experience of 1.5 years. She is an IEEE Member, ACM Member, SIAM Member and Life Member of IAENG. She has more than 50 research papers in peer-reviewed international journals, book chapters, and conferences. One student has been awarded Ph.D. under her supervision and three more are in process. Her research interests include Social Network Analysis, Social Network Mining, Social Media, Data Science, Machine Learning, Data Mining, Web Services, Web Application development and Web Technologies, Software Engineering, Software Testing and Information Retrieval Systems. She is a reviewer of many reputed and peer-reviewed outlets such as IEEE Transactions-TKDE, IEEE Transaction of Cybernetics etc. She is also a reviewer of various Springer, IGI Global, InderScience, and de Gruyter Journals. She has guided more than 15 M.Tech theses and around 100 B.Tech major and minor projects.