Twitter Usage Patterns as a Predictor of User Gender

Parvathy S. Nair
Research Scholar, Department of Media and Communication
Central University of Tamil Nadu, Thiruvarur, +91 8547883203, parvathynair212@gmail.com
&
Francis P. Barclay
Assistant Professor, Department of Media and Communication
Central University of Tamil Nadu, Thiruvarur, +91 7904213765, francis@cutn.ac.in

Abstract— Thanks to the spur of online social networks, users have invented ways of expressing themselves. Though unintended, content posted and their usage patterns, quite often than not, convey pertinent information about social identities and personality traits. Mapping intentions and motivations of Twitter users by studying the content posted by them on the popular microblogging website, the present study aims to build a reliable prediction model for gender. Apart from studying the extent and frequency of usage, a qualitative analysis of tweets was performed to identify usage patterns and able predictors. For the study, Twitter users were chosen and their usage patterns were analysed to spot similarities and differences. Results indicated distinct differences between male and female users with regard to topics discussed and motivations.

Keywords - Twitter, gender studies, social media, new media, discriminant analysis

JM&C
Volume 1 Issue 2
© Central University of Tamil Nadu
Thiruvarur - India

Nair, Parvathy S. and Francis P. Barclay. (2017). Twitter Usage Patterns as a Predictor of User Gender. *Journal of Media and Communication*, 1(2): 1-23. CUTN

Introduction

India is emerging as Twitter's fastest growing market in terms of daily-active users (Chaturvedi, 2017) posting a five-fold market increase than the global average. In January 2005, a survey of social networking websites estimated about 115 million users worldwide. Five years later, Twitter alone accounted for over 200 million users (Yoad, 2015). This popular microblogging website's popularity was rising in India with over 26 million active users in 2017 with an expectation that the count in the largest democracy will cross the 35-million milestone within the next two years (Statista, 2017).

Twitter, along with the other social-networking websites, is changing the sphere of public discourse and setting new trends and agendas, touching upon almost all areas, from the environment and politics to technology and entertainment (Asur, 2010). Twitter is a platform where users present themselves to the world, sometimes revealing personal details and insights into their lives (Golbeck, 2011), sometimes unintentionally (Barclay et al., 2014; 2015; 2015a; 2015b; & 2016). Mining such information could help predict several of their behaviours (Barclay et al., 2017). In the process of creating social-networking profiles, users reveal a lot about themselves through contact details, self-description, status updates, photos and interests (Sun et. al, 2012). The researchers added that text posts represent a large portion of user generated content and contain information which can be used in discovering user attributes. It is, however, easy to provide false name, age, gender and location to hide one's true identity and it would therefore be useful if user profiles can be checked on the basis of available text content. Sun et. al, (2012), further, observed that a common approach of uncovering hidden user attributes in social media is to model writing habits of users by extracting various features from texts they have posted (Golbeck et. al, 2011).

Sumner et al., (2012) devised a model to predict antisocial traits of narcissism and psychopathy. This was performed by comparing the Dark Triad and Big Five personality traits of Twitter users with their profile attributes and use of language and the results shows that there were some statistically significant relationships between these variables.

Eighteenth century logician Augustus De Morgan was the first to suggest that an author could be identified by the characteristics of his or her writing and in this modern time, the person can be identified from the content s/he posts in social media (Zheng et. al, 2006). Park (2015) described a method for assessing personality using an open-vocabulary analysis of language from social media. He compiled the written language from social media users and their questionnaire-based self-reported Big Five personality traits, and built a predictive model of personality based on their language.

According to Ajzen (1987), a multitude of personality traits were identified and new trait dimensions continued to join the growing list. Quercia et al., (2011) set out to analyse the relationship between personality and different types of Twitter users, including popular users and influentials. He gathered personality data, analysed it, and found that both popular users and influentials were extroverts and emotionally stable, and that those popular users were "imaginative" (high in Openness), while influentials tend to be "organised" (high in Conscientiousness). It then showed a way of accurately predicting a user's personality simply based on three counts publicly available on profiles: following, followers, and listed counts.

Personality has been shown to be useful in predicting job satisfaction, professional and romantic relationship success, and even preference for different interfaces (Golbeck et. al, 2011). The researchers presented a method by which a user's personality can be accurately be predicted through the publicly available information on their Twitter profiles. Meanwhile, Hughes (2012) showed that personality was related to online socialising and information seeking and exchange. Quercia (2012), however, found that there was no statistical relationship between Facebook popularity (number of contacts) and personality traits. Though there had been some peripheral examinations of individual differences such as gender and personality, self-presentation had been the primary focus (Muscanell 2012).

Can gender be identified? There is substantial evidence of gender differences in face-to-face communication, and we suspect that similar differences are present in electronic communication (Thomson, 2001). There is a growing interest in predicting the gender and age of authors from texts (Nguyen et al., 2013). Authorship analysis provides a means to glean information about the author of a document originating from the internet or elsewhere, including but not limited to the author's gender (Nguyen et al., 2013). There have been several studies on gender prediction of blogs. Mukherjee (2017) focussed more on the authorial style of both the genders, which are better captured using function words and part of speech n-grams. Goswami et al. (2009) demonstrated that the use of language in blogs correlates with age but could not determine similar correlation with gender.

Large corporations are interested in knowing what types of people (male or female) like their products based on analysis of posts on social media. Peersman et al. (2011) observed that these reviews were helpful in many commercial domains, such as target advertisement and product development. Likewise, intelligence departments may use gender classification for crime investigation. There were a lot of empirical studies devoted to gender specific text analysis and they explored how age and gender affected writing style and topics discussed. Zhang (2010) investigated authorship gender mining from e-mail text documents. Herring (2006) investigated the language/gender/genre relationship in weblogs, a popular new mode of computer-mediated communication (CMC). The task on Author Profiling at PAN 2013 encouraged researchers to identify age and gender of the authors of a large amount of anonymous texts (Rangel et al., 2013).

Establishing demographic data for Twitter users was a key challenge of (Sloan 2015) because Twitter does not make gender or age available. Gender and age are inferred by leveraging profile information, such as gender-discriminating names or crawling for links to publicly available data (Burger et al., 2011). There had been many attempts to profile the demographic characteristics of Twitter users which have drawn on metadata to estimate location, gender, language use (Sloan et al., 2013), occupation, social class, and age (Sloan et al., 2015). Miller et al. (2012) used character level n-grams as a feature to classify Twitter text to predict gender. Data collected from Twitter was often restricted by text length causing further difficulties for gender classification (Miller, 2012). Alowibdi et al. (2013) used non-textual features like background colors and combinations to classify Twitter profiles based on gender and got reasonably high accuracy.

Research Questions

RQ1: Are Gender and Tweet indicators related?
RQ2: What is the relationship among Twitter indicators?
RQ3: Can Gender be predicted using Tweet indicators?

There are well-known linguistic differences between the writing of men and women, and these differences can be effectively used to predict the gender of a document's author (Deitrick et al., 2012).

The study by Bamman et al. (2014) was the only computational study that approached gender as a social variable. The classification of tweets or microblogs by gender was only recently being explored (Mukherjee, 2017). Based on experiments on a subset of the British National Corpus, they found that women have a more relational writing style and men have a more informational writing style (Verhoeven 2017). By clustering Twitter users based on their tweets, they showed that multiple gendered styles existed. Current approaches use supervised machine learning models trained on tweets from males and females (Nguyen et. al, 2013). However, the resulting stereotypical models are ineffective for Twitter users who tweet differently from what is to be expected from their biological sex (Nguyen et. al, 2013).

Many studies have come to similar conclusions regarding which of the features that distinguish male and female authors. It has been reported that women tend to use more emotionally charged language as well as more adjectives and adverbs, and apologise more frequently than men. On the other hand, men tend to use more references to quantity. It has also been observed that gender-specific language was more prevalent in conversations consisting of only one gender when compared to pairs or groups of both genders (Deitrick et al, 2012). Thomson (2001) conducted a discriminant analysis and it showed that it was possible to successfully classify the participants' gender with 91.4 percent accuracy. Hence, in the present study, the emotional content in tweets is analysed to check if they can be used to predict the gender of the Twitter users. Argamon, et al. (2002) found that women tend to use more personal pronouns, whereas male authors use determiners and numbers, along with "he" and "of" more frequently. Herring et al. (2006) then selected a set of male and female preferential specific word forms, following the model of Argamon et al. (2002), to investigate whether gender is a stronger predictor of linguistic variation in weblogs and he found a correlation between language variation and genre. Herring (2006) observed that the diary entries contained more 'female' stylistic features, and the filter entries more 'male' stylistic features.

Schwartz (2013) shed new light on the psychosocial processes and differences among female and male users of social media (males used the possessive 'my' when mentioning their 'wife' or 'girlfriend' more often than females used 'my' with 'husband' or 'boyfriend').

Herring (2004) found a relationship between gender of blog author and blog type: women write more personal journals, while filter-type blogs, albeit a minority overall, are written mostly by men.

All-male groups have been observed to talk more on politics and sports compared to female groups (Coates 1993). In the female category, a topic on food and beverages is also present, but with a different focus (Verhoeven 2017) and male category are associated with drinking and Sports.

Hence, in the present study, going by the observations of Coates (1993), the topics discussed in tweets are analysed to check their power of prediction. Schwartz et al. (2013) and Rangel (2013) showed that words and the language of tweets can be used to predict gender of the user.

Freitas et al. (2015) also had studied patterns of posts and topics to make judgements about Twitter users. Raacke (2008) found that women regularly change various aspects of their profile pages and present a social portrait of themselves. According to Muscanell (2012) women users more frequently post public messages compared to male users. Herring (1996) results noted that academic discussion list attracted female participation and it tended to focus on 'women's' topics, and feminised professions while males predominated in online discussions about politics, philosophy and linguistics. More females than males wrote personal journal blogs (Herring, 2006). The writing in women's blogs was significantly less formal than in men's blogs, as measured by Heylighen (2002). Muscanell (2012) found men using social networking sites for forming new relationships while women reported using them more for relationship maintenance. Males expressed higher negativity and lower desire for social support (Choudhury, 2017). Unlike their male peers, female users in our dataset express more positivity, greater involvement in social and familial concerns.

Hypotheses

H_a1: Tweet indicators (topics and emotions) share an association with Gender of Twitter users.

H_a2: Tweet indicators (topics and emotions) are interrelated.

Based on the strength of the relationships between the tweet indicators and the Gender of the twitter user, a theoretical model could be built to predict the gender of the user.

Research Method

The present study aims to predict Gender of twitter users using Tweet Indicators (emotional content of tweets and the topics discussed in them). Tweet Indicators chosen for the present study are presented in Table 2.1.

Table 2.1 **Table of tweet Indicators**

Topics	Emotions
Personal	Fear
Technology	Anger
Food/Travel	Sadness
Entertainment	Disgust
Fashion	Joyful
Business	Anticipation
Religion	Excitement
Sports	Trust

A sample of 106 Twitter users was chosen using a systematic random sampling procedure. Each of the alphabets were entered in the Twitter Search API and the top users that appeared were chosen for the study. This procedure was repeated till 106 users were chosen.

Expert coders were employed to manually content analyse the tweets and rate them under the categories of Tweet Indicators. Tweets that did not fall under one of the chosen categories were rated as 'others' and excluded from analysis. Apart from the manual content analysis of the tweets of the 106 chosen Twitter users, their Gender was recorded.

Data analysis

Hypothesis testing

Ha1: Tweet indicators (topics and emotions) share an association with Gender of Twitter users

To test the relationships between Gender and the Tweet Indicators, choosing 106 active Twitter users, T-Test was performed and the test results are presented in Table 3.1 and Fig. 3.1.

Table 3.1 **T-Test Results: Gender vs. Tweet Indicators**

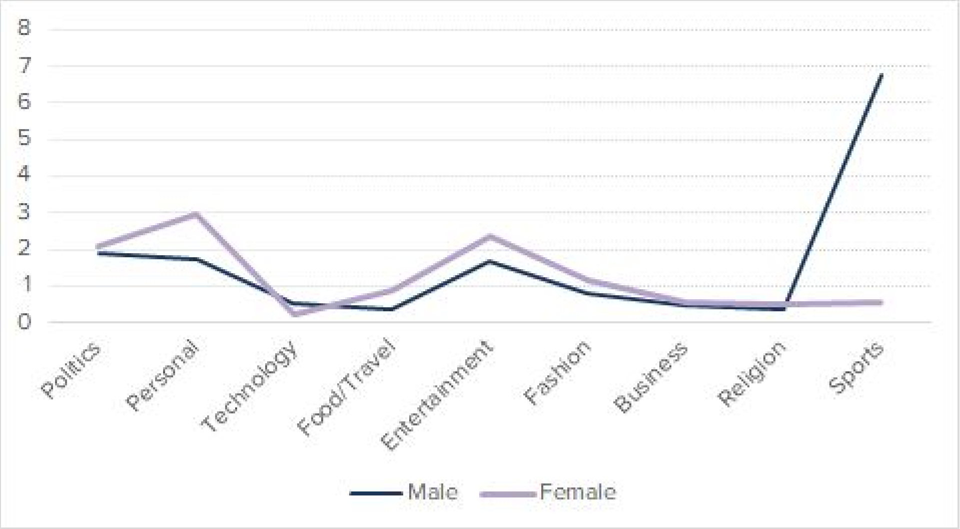
		Gender	Mean	Std. Error Mean	Sig. Value
TOPICS	Politics	Male	1.89	.473	.647
		Female	2.06	.464	
	Personal	Male	1.75	.200	0.005
		Female	2.98	.344	
	Tech	Male	.55	.123	0.000
		Female	.23	.056	
	Food / travel	Male	.36	.122	0.019
		Female	.88	.272	
	Entertainment	Male	1.66	.259	0.110
		Female	2.38	.572	
	Fashion	Male	.78	.338	0.634
		Female	1.13	.268	
	Business	Male	.46	.100	0.163
		Female	.55	.191	
	Religion	Male	.39	.089	0.145
		Female	.50	.147	
	Sports	Male	6.76	2.606	0.000
		Female	.56	.128	
EMOTIONS	Fear	Male	.38	.134	0.058
		Female	.71	.199	
	Anger	Male	.90	.270	0.035
		Female	1.58	.374	
	Sadness	Male	1.44	.328	0.150
		Female	2.06	.414	
	Disgust	Male	.82	.237	0.027
		Female	.44	.112	
	Joy	Male	4.59	.939	0.348

		Female	6.03	1.299	
	Anticipation	Male	1.39	.249	0.936
		Female	1.29	.348	
	Excitement	Male	2.62	.451	0.411
		Female	2.30	.406	
	Trust	Male	.62	.144	0.232
		Female	.98	.237	

T-Test results indicated a statistically-significant association between Gender and both Personal (*F*=7.930, *p*< .0005) and Technology (F=16.9, p< .0005). It can be inferred that female Twitter users shared more tweets on personal issues (mean = 2.98) compared to male users (mean = 1.75).

On the other hand, male Twitter users shared more information on Technology (mean = .55) than their female counterparts (mean = .23).

Fig. 3.1 **Plot: Gender vs. Tweet Indicators**



T-Test results also indicated a statistically-significant association between Gender and Food / Travel (*F*= 5.549, *p* = .019). It can be inferred that female Twitter users share more posts related to Food / Travel (mean =.88) compared to their male counterparts (mean =.36). T-Test results indicated a statistically-significant association between Gender and the dependent variables Sports (F = 16.87, p < .0005). Male Twitter users shared more on Sports (mean = 6.76) compared to female users (mean = .56). Similarly, Gender was related to Anger (F = 4.415, p = .35) and Disgust (F = 4.925, p = .27). Female users (mean = 1.58) shared more Anger on Twitter compared to men (mean = .90). On the contrary, male users shared more Disgust (mean = .82) compared to their female counterparts (mean = .44).

Hypothesis testing
Ha2: Tweet indicators (topics and emotions) are interrelated

To test the relationships among the Topic variables, a Pearson correlation test was run and the results are presented in Table 3.2.

Table 3.2 **Correlation results (part 1)**

		Personal	Tech	Food	Ent.	Fashion	Biz	Religion	Sports
Politics	Corr	.216**	.320 *	.161**	.228**	.546**	.366*	.557**	.569**
	Sig.	.001	.000	.008	.000	.000	.000	.000	.000
Pers.	Corr		-.165	.419**	.204**	.073	.211*	.408**	.043
	Sig		.007	.000	.001	.139	.001	.000	.263
Tech	Corr			.045	-.037	.301**	.151*	.114*	.362**
	Sig.			.249	.290	.000	.012	.044	.000
Food/ Travel	Corr				.154*	.366**	.002	.124*	.268**
	Sig.				.011	.000	.490	.032	.000
Ent.	Corr					.159**	.265*	.179**	.019
	Sig					.008	.000	.004	.386
Fashion	Corr						.291*	.212**	.747**
	Sig.						.000	.001	.000
Biz.	Corr							.405**	.033
	Sig.							.000	.313
Religion	Corr								.209**
	Sig.								.001

***. Correlation is significant at the 0.01 level (1-tailed).*
**. Correlation is significant at the 0.05 level (1-tailed).*

The data showed no violation of normality, linearity and homoscedasticity. There was a strong, positive and statistically-significant correlation between the number of Political tweets the Twitter users had posted and the number of Personal tweets they posted (r = .216, n = 106, p = .001); between the number of Political tweets the Twitter users had posted and the number of Tech tweets they posted (r = 320, n = 106, p< .0005).

There was a positive and statistically-significant correlation between the number of Political tweets the Twitter user had posted and the number of Food/ Travel tweets they posted (r = .161, n = 106, p = .008); between the number of Political tweets the Twitter user had posted and the number of Entertainment tweets they posted (r = .228, n = 106, p < .005); between the number of Political tweets the Twitter user had posted and the number of Fashion tweets they posted (r = .546, n=106, p<.0005); between the number of Political tweets the Twitter user had posted and the number of Business tweets they posted (r = .366, n = 106, p <. 0005); between the number of Political tweets the Twitter user had posted and the number of Religious tweets they posted (r = .557, n = 106, p < .0005); between the number of Political tweets the Twitter user had posted and the number of Sports tweets they posted (r = .569, n = 106, p <.0005). Similarly, there was a correlation between the number of Personal tweets the Twitter users had posted and the number of Tech tweets they posted (r= -.165, n = 106, p = .007); between the number of Personal tweets and the number of Food/ Travel tweets they posted (r = .419, n = 106, p = .021). There was a positive and statistically-significant correlation between the number of Personal tweets the Twitter user had posted and the number of Entertainment tweets they posted (r = .204, n = 106, p = .001); between the number of Personal tweets the Twitter user had posted and the number of Business tweets they posted (r = .211, n = 106, p = .001); between the number of Personal tweets the Twitter user had posted and the number of Religious tweets (r= .408, n = 106, p < .0005); However, there was no statistically-significant relationship between the number of Personal tweets, and Fashion and Sports tweets.

There was a positive and statistically-significant correlation between the number of Tech tweets Twitter user had posted and the number of

Fashion tweets they posted (r = .301, n = 106, p < .0005); between the number of Tech tweets Twitter user had posted and the number of Sports tweets they posted (r = .362, n = 106, p <.0005). However, there was no statistically-significant relationship between the number of Tech tweets and the Food/ Travel, Entertainment, Religion and Business tweets posted. There was a strong, positive and statistically-significant correlation between the number of Food/ Travel tweets Twitter user had posted and the number of Fashion tweets they posted (r = .366, n = 106, p < .0005); between the number of Food/ Travel tweets Twitter user had posted and the number of Sports tweets they posted (r = .268, n = 106, p < .0005).

There was a positive and statistically-significant correlation between the number of Entertainment tweets Twitter user had posted and the number of Fashion tweets they posted (r = .159, n = 106, p = .008); between the number of tweets Entertainment Twitter user had posted and the number of Business tweets they posted (r = .265, n = 105, p <.0005); between the number of tweets Entertainment Twitter user had posted and the number of Religious tweets they posted (r = .179, n = 106, p = .004). However, there was no statistically-significant relationship between the number of Entertainment tweets and and Sports tweets posted. There was a positive and statistically-significant correlation between the number of Fashion tweets Twitter user had posted and the number of Business tweets they posted (r = .291, n = 106, p <.0005); between the number of Fashion tweets Twitter user had posted and the number of Religious tweets they posted (r = .212, n = 106, p =.001); between the number of Fashion tweets Twitter user had posted and the number of Sports tweets they posted (r = .747, n = 106, p < .0005). Similarly, there was a strong, positive and statistically-significant correlation between the number of Business tweets Twitter user had posted and the number of Religious tweets they posted (r = .405, n = 105, p < .0005).

There was a positive and statistically-significant correlation between the number of Religious tweets Twitter user had posted and the number of Sports tweets they posted (r = .209, n = 106, p = .001). To test the relationships among the Emotion variables, a Pearson product-moment correlation test was run and the results are presented in Tables 3.3. There was a positive and statistically-significant correlation between the number of Fear tweets Twitter user had posted and the number of Anger tweets they posted (r = .652, n = 106, p < .0005); between the number of Fear tweets Twitter user had posted and the number of Sadness tweets they posted (r = .716, n = 106, p <

.0005); between the number of Fear tweets Twitter user had posted and the number of Joyful tweets they posted (r = .576, n = 106, p < .0005); between the number of Fear tweets Twitter user had posted and the number of Anticipation tweets they posted (r = .701, n = 106, p < .0005); between the number of Fear tweets Twitter user had posted and the number of Excitement tweets they posted (r= .659, n = 106, p < .0005); between the number of Fear tweets Twitter user had posted and the number of Trust tweets they posted (r = .725, n = 106, p < .0005).

Table 3.3 **Correlation results (part 2)**

		Anger	Sadness	Disgust	Joy	Anticipation	Exit	Trust
Fear	Corr.	.652**	.716**	.576**	.746**	.701**	.659**	.725**
	Sig.	.000	.000	.000	.000	.000	.000	.000
Anger	Corr.		.860**	.764**	.724**	.273**	.629**	.464**
	Sig.		.000	.000	.000	.000	.000	.000
Sadness	Corr.			.693**	.881**	.353**	.798**	.641**
	Sig.			.000	.000	.000	.000	.000
Disgust	Corr.				.560**	.476**	.639**	.304**
	Sig.				.000	.000	.000	.000
Joy	Corr.					.406**	.904**	.753**
	Sig.					.000	.000	.000
Anticipation	Corr.						.462**	.463**
	Sig.						.000	.000
Excitement	Corr.							.611**
	Sig.							.000

There was a strong, positive and statistically-significant correlation between the number of Anger tweets Twitter user had posted and the number of Sadness tweets they posted (r= .860, n = 106, p < .0005); between the number of Anger tweets Twitter user had posted and the number of

Disgust tweets they posted (*r* = .764, n = 106, *p* < .0005); between the number of Anger tweets Twitter user had posted and the number of Joyful tweets they posted (*r* =.724., n = 106, *p* < .0005); between the number of Anger tweets Twitter user had posted and the number of Anticipation tweets they posted (*r* =.273., n = 106, *p* <.0005); between the number of Anger tweets Twitter user had posted and the number of Excitement tweets they posted (*r*= .629, n = 106, *p* < .0005); between the number of Anger tweets Twitter user had posted and the number of Trust tweets they posted (*r* = .464., n = 106, *p* < .0005).

There was a strong, positive and statistically-significant correlation between the number of Sadness tweets Twitter user had posted and the number of Disgust tweets they posted (*r* = .693, n = 106, *p* < .0005); between the number of Sadness tweets Twitter user had posted and the number of Joyful tweets they posted (*r* = .881, n = 106, *p* < .0005); between the number of Sadness tweets Twitter user had posted and the number of Anticipation tweets they posted (*r* = .353, n = 106, *p* < .0005); between the number of Sadness tweets Twitter user had posted and the number of Excitement tweets they posted (*r* = .798, n = 106, *p* < .0005); between the number of Sadness tweets Twitter user had posted and the number of Trust tweets they posted (*r* = .641, n = 106, *p* < .0005). It can be inferred that the more the number of Sadness tweets, the more were the numbers of Disgust, Joyful, Anticipation, Excitement and Trust tweets.

There was a strong, positive and statistically-significant correlation between the number of Disgust tweets Twitter user had posted and the number of Joyful tweets they posted (*r* = .560, n = 106, *p* < .0005); between the number of Disgust tweets Twitter user had posted and the number of Anticipation tweets they posted (*r* = .476, n = 106, *p* < .0005); between the number of Disgust tweets Twitter user had posted and the number of Excitement tweets they posted (*r* = .639, n = 106, *p* < .0005); between the number of Disgust tweets Twitter user had posted and the number of Trust tweets they posted (*r* = .304, n = 106, *p* < .0005). There was a strong, positive and statistically-significant correlation between the number of Joyful tweets Twitter user had posted and the number of Anticipation tweets they posted (*r* = .406, n = 106, *p* <.0005); between the number of Joyful tweets Twitter user had posted and the number of Excitement tweets they posted (*r* = .904, n = 106, *p* < .0005); between the number of Joyful tweets Twitter user had posted and the number of Trust tweets they posted (*r* = .753, n = 106, *p* < .0005). Similarly, there was a strong, positive and statistically-significant correlation between the number of Anticipation tweets Twitter user had posted and the number of Excitement tweets they posted ($r = .462$, n = 106, $p < .0005$); between the number of Anticipation tweets Twitter user had posted and the number of Trust tweets they posted ($r = .463$, n =106, $p < .0005$).

There was a strong, positive and statistically-significant correlation between the number of Excitement tweets Twitter user had posted and the number of Trust tweets they posted ($r = .611$, n = 106, $p < .0005$). To test the predictive abilities of the Tweet indicators, a Discriminant analysis was performed with Gender as the outcome variable. For the variable Gender, there were two groups: Male and Female.

Finally, seven independent variables were chosen for Discriminant analysis to predict Gender, using Structure Matrix. The discriminant analysis was used to build a prediction model for the variable Gender using the Twitter variables and the test results are presented in Table 3.6

Table 3.6 **Discriminant analysis for Gender**

	Wilks' Lambda	F	df1	df2	Sig.
Personal	0.956	10.223	1	223	0.002
Tech	0.979	4.852	1	223	0.029
Sports	0.979	4.765	1	223	0.03
Food/Travel	0.985	3.416	1	223	0.066
Anger	0.99	2.27	1	223	0.133
Fear	0.991	2.015	1	223	0.157
Disgust	0.992	1.902	1	223	0.169

Log Determinants

Gender	Rank	Log Determinant
Male	10	34.321
Female	10	41.869
Pooled within-groups	10	46.597

Ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		1967.645
F	Approx.	34.076
	df1	55
	df2	151563.293
	Sig.	0.000

Tests null hypothesis of equal population covariance matrices.
Both Box's M score and the Wilks' Lambda indicated that the model was acceptable.

Canonical Correlation coefficient was .399, indicating that the independent variables explained about 16 percent (r^2) of variation in the dependent variable.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.190a	100.0	100.0	.399

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.840	37.882	10	.000

Standardized Canonical Discriminant Function Coefficients

	Function (1)
Personal	0.332
Tech	-0.07
Sports	-0.217
Food / Travel	-0.181
Anger	-0.005
Fear	0.582
Disgust	-0.574

Structure Matrix

	Function (1)
Personal	0.491
Tech	-0.339
Sports	-0.336
Food/Travel	0.284
Anger	0.232
Fear	0.218
Disgust	-0.212

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions. Variables ordered by absolute size of correlation within function.

Reviewing the Structure Matrix, it can be inferred that the number of Personal tweets posted by an user, Tech-related, Sports and Food and Travels tweets were the leading indicators for the prediction of the Gender of the user. Among the emotions expressed on Twitter, Anger, Fear and Disgust turned out to be useful indicators for Gender of the Twitter user. Though the effect sizes were minimal, these indicators exhibit a potential to build prediction models for gender as well as other personal details.

Functions at Group Centroids

Gender	Function (1)
Male	-.398
Female	.472

Unstandardized canonical discriminant functions evaluated at group means

Classification Function Coefficients

	Male	Female
Personal	.325	.426
Tech	.593	.536
Sports	.002	-.007
Food / Travel	-.158	-.233
Anger	-.196	-.197
Fear	.140	.430
Disgust	.151	-.090
(Constant)	-1.063	-1.727

Fisher's linear discriminant functions

Classification Results

Gender			Predicted Group Membership		Total
			Male	Female	
Original	Count	Male	111	11	122
		Female	63	40	103
	%	Male	91.0	9.0	100.0
		Female	61.2	38.8	100.0

a. 67.1% of original grouped cases correctly classified.

It also can be inferred that female users had considerably higher number of followers and following counts, compared to their male counterparts. Similarly, women posted more Personal tweets, compared to men. Male users tweeted more on Food/ Travel, Technology and Sports compared to their female counterparts. Female users expressed more Fear, compared to their male counterparts.

Men expressed more Anger and Disgust than women. And, the prediction model for Gender as determined by the Discriminant Analysis is:

Gender (1, 2) = **.332 (Personal) - .070 (Tech) - .217 (Sports) -.181 (Food / Travel) -.005 (Anger) + .582 (Fear) -.574 (Disgust)**

Discussion

Study results have shown that several of the Tweet Indicators were related to Gender of the Twitter users and exhibited the potential to predict sex. Further, correlation tests revealed that these Tweet Indicators were interrelated, demanding factorial analysis to test their relationships with the variable Gender. Discriminant analysis was used to build a valid prediction model for Gender of the Twitter users using simple topics and emotions of tweets. As prophesied by Zhenget. al. (2006), Mukherjee (2017), Asur (2010) and Schwartz (2013), tweets could be an important tool to estimate hidden user attributes, as supported by the results of the present study. Findings of the present study also align with the thoughts of Golbeck (2011) and Summer et al., (2012) who observed that tweets could divulge personal details and insights into the lives of the users. In line with the findings of Nguyen et al., (2013), Deitrick et al., (2012) Argamon, et al. (2002), the present study has found that Tweet Indicators could be able predictors of Gender of users. Verhoeven (2017) and Heylighen (2002) differentiated the writing style of men and women. The present study result also matches with the same.

Conclusion

As the study results indicate, Tweet Indicators could be used to predict the Gender of a Twitter user. Several of the topical and emotional Tweet Indicators—Personal, Food/Travel, Sports, Technology, Anger, Fear and Disgust—were found to have significant relationships with Gender of the

Twitter users. Discriminant analysis was used to build a prediction model for the Gender of Twitter users using publicly-available data.

References

Ajzen, I. (1987). Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. Advances in experimental social psychology, 20, 1-63.

Alowibdi, J. S., Buy, U. A., & Yu, P. (2013). Language independent gender classification on Twitter. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 739-743). ACM.

Argamon, S., Koppel, M., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412.

Asur Sitaram, Bernardo A. Huberman (2010) Predicting the Future with Social Media. Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference. IEEE, retrieved Oct 24 from

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. Journal of Sociolinguistics, 18(2), 135-160.

Barclay, Francis P. (2014). Political opinion expressed in social media and election outcomes-US presidential elections 2012. GSTF Journal on Media & Communications (JMC), 1(2).

Barclay, Francis P., Pichandy, C., Venkat, A. and Sudhakaran, S. (2015). India 2014: Facebook 'like' as a predictor of election outcomes. Asian Journal of Political Science, 23(2): 134-160.

Barclay, F. P. (2015a). Inter-Media Interaction and Effects in An Integrated Model of Political Communication: India 2014. Global Media Journal, 13(25).

Barclay, F. P., Pichandy, C., & Venkat, A. (2015b). India elections 2014: Time-lagged correlation between media bias and facebook trend. Global Journal of Human-Social Science Research.

Barclay, F. P., Pichandy, C., Venkat, A., & Sudhakaran, S. (2016). Twitter Sentiments: Pattern Recognition and Poll Prediction. In Communication and Information Technologies Annual: [New] Media Cultures (pp. 141-167). Emerald Group Publishing Limited.

Barclay, Francis P. (2017). Media effect on media: Progression of political news and tweets during India 2014. Journal of Media and Communication, 1(1): 1-28. CUTN.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301-1309). Association for Computational Linguistics.

Chaturvedi, Anumeha.(2017). How India emerged as Twitter's fastest growing market in terms of daily active users. The Economic Times, published on May 6, retrieved Oct 19, from https://economictimes.indiatimes.com/opinion/interviews/india-became-our-number-one-market-in-daily-users-twitters-new-india-director-taranjeet-singh/articleshow/58601906.cms

Coates, Jennifer. (1993). Women, Men, and Language (2nd edition). London: Longman.

De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., & Nielsen, R. C. (2017, February).

Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In CSCW (pp. 353-369).

Deitrick William, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, Wei Hu(2012). Author Gender Prediction in an Email Stream Using Neural Networks. Journal of Intelligent Learning Systems and Applications, 2012, 4, 169-175.

Freitas, C., Benevenuto, F., Ghosh, S., & Veloso, A. (2015). Reverse engineering socialbot infiltration strategies in twitter. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 25-32). ACM.

Golbeck Jennifer; Cristina Robles; Michon Edmondson; Karen Turner (2011).Predicting Personality from Twitter. Privacy, Security, Risk and Trust (PASSAT) and IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference. IEEE

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In Third International AAAI Conference on Weblogs and Social Media.

Herring, S. C. (2004). Computer-mediated Communication. Language and woman's place: Text and commentaries, 216.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. Journal of Sociolinguistics, 10(4), 439-459.

Heylighen, Francis & Jean-Marc Dewaele. (2002). Variation in the contextuality of language: An empirical measure. Foundations of Science 6: 293–340.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. Computers in Human Behavior, 28(2), 561-569.

Miller, Z., Dickinson, B., & Hu, W.(2012). Gender prediction on twitter using stream algorithms with n-gram character features. International Journal of Intelligence Science, 2(04), 143.

Mukherjee, S., & Bala, P. K. (2017). Gender classification of microblog text based on authorial style. Information Systems and e-Business Management, 15(1), 117-138.

Muscanell, N. L., & Guadagno, R. E. (2012). Make new friends or keep the old: Gender and personality differences in social networking use. Computers in Human Behavior, 28(1), 107-112.

Nguyen Dong, DolfTrieschnigg, Seza Dogru, Rilana Grave, Mariet Theune, Theo Meder, Franciska de Jong. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pages 1950–1961, retrieved Oct 30 from http://www.aclweb.org/anthology/C14-1184

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J & Seligman, M. E.(2015). Automatic personality assessment through social media language. Journal of personality and social psychology, 108(6), 934.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L.(2011). Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J.(2011, October). Our twitter profiles, our selves: Predicting personality with twitter. In Privacy, Security, Risk and Trust

(PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 180-185). IEEE.

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., & Crowcroft, J. (2012, February). The personality of popular facebook users. In Proceedings of the ACM 2012 conference on computer supported cooperative work (pp. 955-964). ACM.

Rangel, F., & Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. Natural Language Processing and Cognitive Science, 177.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at PAN 2013. Notebook Papers of CLEF, 23-26. Retrieved on Oct 30 from http://dx.doi.org/10.4236/jilsa.2012.43017 Published Online August 2012

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M. & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. PloS one, 10(3), e0115545.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., &Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. Sociological research online, 18(3), 7.

Statista.(2017). Number of Twitter users in India from 2012 to 2019 (in millions). Statista, retrieved Oct 19, from https://www.statista.com/statistics/381832/twitter-users-india/

Sumner, C., Byers, A., Boochever, R., and Park, G. J.(2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In Machine learning and applications (ICMLA), 2012 11th international conference, 2: 386-393. IEEE.

Sun, X., Ding, X., & Liu, T. (2014). Gender Identification on Social Media. In Chinese National Conference on Social Media Processing (pp. 99-107). Springer, Berlin, Heidelberg, retrieved Oct 24 from

Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. British Journal of Social Psychology, 40(2), 193-208.

Verhoeven, B., Škrjanec, I., & Pollak, S. (2017). Gender Profiling for Slovene Twitter Communication: The Influence of Gender Marking, Content and Style. BSNLP 2017, 119.

Yoad Lewenberg, Yoram Bachrach, Svitlana Volkova. (2015). Using emotions to predict user interest areas in online social networks. Data Science and Advanced Analytics (DSAA) 2015. 36678 2015. IEEE International Conference on, pp. 1-10, 2015.

Zhang, C., & Zhang, P. (2010). Predicting gender from blog posts. University of Massachusetts Amherst, USA.

Zheng, J. Li, H. Chen and Z. Huang, “A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques,” Journal of the American Society for Information Science and Technology, Vol. 57, No. 3, 2006, pp. 378-393.

Parvathy S. Nair is a postgraduate student in Department of Media and Communication, School of Communication, Central University of Tamil Nadu, Thiruvarur, India.

Francis P. Barclay is Assistant Professor in the Department of Media and Communication, School of Communication, Central University of Tamil Nadu, India. Dr. Barclay is also a journalist, writer, psephologist and media researcher. He has published and contributed chapters to several books, apart from research articles in reputed journals. His research area is media and politics. He has served several English newspapers in India. His works are available at http://www.francisbarclay.com.