Social Sciences & Humanities Open 4 (2021) 100178

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Comparison of machine learning algorithms for content based personality resolution of tweets

Shruti Garg [a,*], Ashwani Garg [b]

[a] Birla Institute of Technology, Mesra, Ranchi, 835215, India
[b] Faculty of Health, Biomedical Science and Medical Science, Griffith University, Queensland, 4122, Australia

ARTICLE INFO

Keywords:
Machine learning (ML)
MBTI
BIG5
Twitter
Personality resolution

ABSTRACT

The content of social media (SM) is expanding quickly with individuals sharing their feelings in a variety of ways, all of which depict their personalities to varying degrees. This study endeavored to build a system that could predict an individual's personality through SM conversation. Four BIG5 personality items (i.e. Extraversion (EXT), Consciousness (CON), Agreeable (AGR) and Openness to Experiences (OPN) equivalent to the Myers–-Briggs Type Indicator (MBTI)) were predicted using six supervised machine learning (SML) algorithms. In order to handle unstructured and unbalanced SM conversations, three feature extraction methods (i.e. term frequency and inverse document frequency (TF-IDF), the bag of words (BOW) and the global vector for word representation (GloVe)) were used. The TF-IDF method of feature extraction produces 2–9% higher accuracy than word2vec representation. GloVe is advocated as a better feature extractor because it maintains the spatial information of words.

1. Introduction

Different social media (SM) platforms are increasing their user numbers, including 917 million visitors a month on LinkedIn, 3.62 billion visitors on Twitter, 22.77 billion visitors on YouTube and 2.86 billion visitors on Instagram (Patel, 2020). Moreover, the volume of text posts on each site is increasing by 20–30% daily (Patel, 2020). Consequently, social networking has become the most widely utilized correspondence and association instrument between individuals over the past years, especially during the spread of COVID-19 (Ghareb et al., 2018).

SM has been analyzed by many researchers through natural language processing because its content is very much unstructured (Louis, 2016). Sentiment analysis is the most popular research that has been conducted through SM content (Kharde & Sonawane, 2016). However, sentiment analysis is limited by its two-way outcome, which is not an efficient means to define real-life scenarios (Acosta et al., 2017). For example, saying a person is positive or negative does not adequately define their personality. Also, the content of SM is so vast that it is impossible to process without machines, a comparison presented in (Enos et al., 2006).

The personality of a person is defined by the typical sets of behaviors, perceptions and sentimental feelings that evolve from human and atmospheric factors (Hogan et al., 1991). There are many personality models available, including the Big Five (BIG5) personality traits (Roccas et al., 2002), the Myers–Briggs Type Indicator (MBTI) (Truity, 2020) and Dominance, Influence, Steadiness, Compliance (DISC) assessment (Ahmad and Siddique, 2017). MBTI and BIG5 were found to be widely used in the literature (Celli & Lepri, 2018). The MBTI personality indicators could be related to the BIG5 personality terms given in (Furnham et al., 2003) as shown in Table 1.

An equivalent BIG5 personality model is used for prediction in the present work because the dataset is in the form of an MBTI personality indicator.

Several personality assessment tests are available on the internet, including BIG5, MBTI, Enneagram, career test and DISC assessment (Truity, 2020). These tests are helpful for the assessment of personalities. However, a major disadvantage of online tests is that they are very time-consuming and cannot be regarded as scientific assessments (Patrick, 2020). SM is a platform on which people share their actual opinions (Kietzmann & Canhoto, 2013). Therefore, the present research has adopted supervised machine learning (SML) to identify personalities from SM posts using the BIG5 personality model.

The subsequent sections of this paper describe related studies in Section 2 and methodology in Section 3, while results are described in

* Corresponding author.
E-mail address: gshruti@bitmesra.ac.in (S. Garg).

https://doi.org/10.1016/j.ssaho.2021.100178
Received 8 June 2020; Received in revised form 5 April 2021; Accepted 18 June 2021
Available online 1 July 2021

Table 1
Correlation of MBTI personality indicator to BIG5 personality items.

MBTI Personality Indicators	BIG5 Personality Items
Introversion vs. Extraversion (I–E)	Extraversion (EXT)
Intuition vs. Sensing (N–S)	Openness (OPN)
Thinking vs. Feeling (T–F)	Agreeableness (AGR)
Judging vs. Perceiving (J–P)	Consciousness (CON)
*Neuroticism cannot be mapped with MBTI indicator	

Section 4 and discussion and conclusions are stated in Section 5.

2. Related studies

Personality assessment has been found to be a helpful tool by many user groups, such as human resource (HR) managers, psychiatrists, intelligence agencies, family members and friends. Different domains wherein personalities are assessed are stated in (Pesic et al., 2019). Some of the situations in which personalities were assessed in recent years are shown in Table 2.

Email-based personality identification at the workplace was conducted in Ezpeleta et al. (2020) and Russell & Woods, (2020). Military applications for generating threats or spreading unauthorized information in cyber domains were developed in (Sartonen et al., 2020). The target audience listening to these personalities was also included in this work. A combination of enhanced extended nearest neighbor (EENN) and particle swarm optimization (PSO) has been used to identify two types of personality (i.e. openness to experience and extraversion in online learning). The accuracy of this method was found to be 97.6% among all machine learning algorithms (Sun, Wu, & Xiao, 2019).

Mood and personality studies of 130 mobile users have been conducted using signature-based machine learning models (Arribas et al., 2018). The accuracy achieved was 75% for all three classes. Personalities were identified by handwriting in (Thomas et al., 2020) and a correlation of individuals' financial behavior with their personalities was established in cited work.

Two personality types, designated as extraversion and neuroticism, have been identified by two types of datasets, namely the Human Activity dataset and the Ten Item Personality Measure (TIPI) questionnaire, and an association between habits and personality scores has been established. The correlation is observed for the same user, and the activity trait is classified within the same cluster group with which the personality of the user is associated (Lee & Bastos, 2020).

Table 2
Studies on personality assessment.

Ref. No.	Platform	Dataset	Personality Model	Outcome	Methods
(Ezpeleta, Velez de Mendizabal, Hidalgo, & Zurutuza, 2020)	Emails	CSDMC 2010 Spam corpus	MBTI	Sentiment and Personality identification	Discriminative Multinomial, Bayesian Classifier
(Russell & Woods, 2020)	Work Emails	Interview Based	BIG5	Five factors of personality	Thematic analysis of data
(Sartonen et al., 2020)	Cyber Domain	Cyber Domain	No Model	Binary	Target Audience Analysis within the framework of psychological operations (PSYOPS)
(Sun, Wu, & Xiao, 2019)	Online learning	Questionnaire	BIG5	Openness to experience and extraversion	Enhanced Extended nearest neighbor (EENN) with PSO
(Arribas et al., 2018)	Mobile Technologies	Data Collected from wearable devices combination with mobile app	Borderline Personality Disorder	Multiclass 3 class 1. Bipolar mood disorder 2. Borderline Personality disorder 3. Healthy	Signature Based learning
(Thomas et al., 2020)	Handwriting	Handwriting	BIG5	Correlation between handwriting to personality and financial behaviour	Convolution Neural Network
(Lee & Bastos, 2020)	Mobile Technologies	Human Activity Dataset	BIG5	Extroversion and Neuroticism	Correlation and Clustering

Table 3
Studies on personality assessment in SM.

Ref. No.	Platform	Dataset	Personality Model	Outcome	Methods
(Sumner et al., 2012)	Twitter	2927 Twitter users from 89 countries	Dark traits and BIG5	Relationships between Twitter activity, Dark Triad and Big Five personality traits,	Supervised Machine Learning
(Ortigosa et al., 2014)	Facebook	TP2010 Spanish questionnaire	Zuckerman-Kuhlman Personality Questionnaire ZKPQ-50	Multiclass 5 class 1. Activity 2. Aggression Hostility 3. Sociability 4. Impulsive sensational seeking 5. Neuroticism-Anxiety	Weka data mining tool
(Ahmad and Siddique, 2017)	Twitter	Tweets	DISC	Multiclass 4 class 1. Dominance 2. Influence 3. Steadiness 4. Compliance	Natural Language Processing
(Moreno et al., 2019)	Twitter	PAN CLEF 2015	BIG5	Multiclass 5 personalities	SVM, LR and RF with (PCA), (LDA) and Non-(NMF)
(Sengupta & Ghosh, 2020)	Twitter, Facebook, LinkedIn, Trumlr	Real Time	Semantic analysis	Binary	Ensemble learning
(Gjurković et al., 2020)	Redditt	Pandora	BIG5, MBTI	4 -way classification	Regression, Deep learning

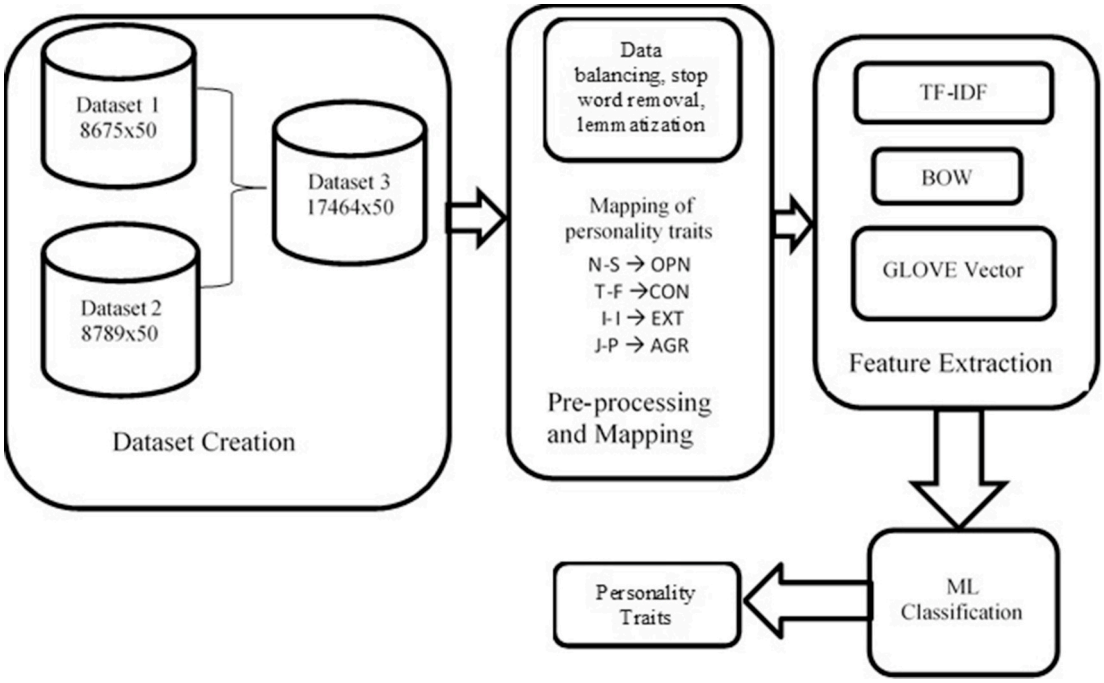

Fig. 1. Workflow of present research.

2.1. Personality assessment on SM

Table 3 shows several pieces of research into text-based personality identification on SM platforms.

Machine learning was used to conclude a significant correlation between Dark Triad personality traits and activity on Twitter in (Sumner et al., 2012). Personalities were identified in five classes: impulsive sensation seeking, neuroticism-anxiety, aggression-hostility, sociability and activity, using data mining algorithms in Weka in (Ortigosa et al., 2014). Personalities were identified for the DISC model using tweets in (Ahmad and Siddique, 2017). The tweets were pre-processed using a natural language processing module of R and divided into four categories: dominance, influence, steadiness and compliance, based on the frequency of tag words in the work.

BIG5 personality traits of Twitter users have been identified using linear support vector classifiers, logistic regression and random forest in (Moreno et al., 2019). Features were extracted from principal components analysis (PCA), linear discriminant analysis (LDA) and non-negative matrix factorization (NMF) before applying machine learning algorithms. Emphasis is given to feature extraction in this.

Real-time semantic analysis using ensemble learning was conducted on Twitter and features were extracted in real time. An accuracy of 79.70% was achieved by logistic regression (Sengupta & Ghosh, 2020). Natural language processing was implemented to find three types of personality models in the PANDORA dataset consisting of 10,000 Reddit user comments. Six regression models (i.e. age and BIG5 personality traits) and eight classification models comprising four dimensions of MBTI, gender and region. The Enneagram was predicted using linear/logistic regression (LR) and deep learning (DL). The deep learning model performed poorly because of the large number of comments per user in the PANDORA dataset (Gjurković et al., 2020).

Many researchers have identified personalities by applying different personality models to SM (Bleidorn & Hopwood, 2019; Kaushal & Patwardhan, 2018). Of all the SM platforms, Twitter has become the most popular among those researching personality identification (Bhavya et al., 2020). Consequently, the personalities of bloggers have been identified here through messages using supervised machine learning techniques.

3. Methodology

People are more involved in virtual than physical socializing during this era of COVID-19 (Boberg et al., 2020). Facebook, Instagram and Twitter are the most widely used SM platforms in the world (Kircaburun et al., 2020). Twitter is one example of an SM platform that provides blogging and instant messaging facilities. The messages on Twitter are available in the form of blogs, tweets, status updates and retweets. Users share short messages with a maximum of 140 characters, which are known as tweets. The words available in these tweets were found to be important in defining the contents (Burnap & Williams, 2015; Srivastava & Roychoudhury, 2020) that are used to identify the personalities of these users.

The contents of each tweet, such as words, music, embedded URLs, images, question and exclamation marks and ellipses, were used to classify the tweets according to the BIG5 personality model using machine learning algorithms.

The six machine learning classifiers were applied here in Python: support vector machine (SVM) (Noble, 2006), decision tree (DT) (Safavian & Landgrebe, 1991), k-nearest neighbors (KNN) (Guo et al., 2003), logistic regression (LR) (Kleinbaum et al., 2002), random forest (RF) (Vijay et al., 2020) and extreme gradient boosting (XGB) (Kunte & Panicker, 2020).

Steps to identify personalities from the content of their tweets are as follows: (i) Data collection (ii) Pre-processing and mapping of data collected (iii) Feature extraction and (iv) Classification using machine learning methods. The execution of the present research is shown in Fig. 1.

3.1. Dataset description

The dataset used in this research was created using two datasets. Dataset1 consists of 8675 users with 50 tweets per user taken from the internet (Mitchell, 2017), and Dataset2 consists of 8789 users with 50 tweets per user collected from Twitter application programming interfaces (API). The two datasets were merged and balanced for personality resolution. The final dataset comprised 17,464 users with 50 tweets per user. Self-declared tweets were collected from Twitter's API and

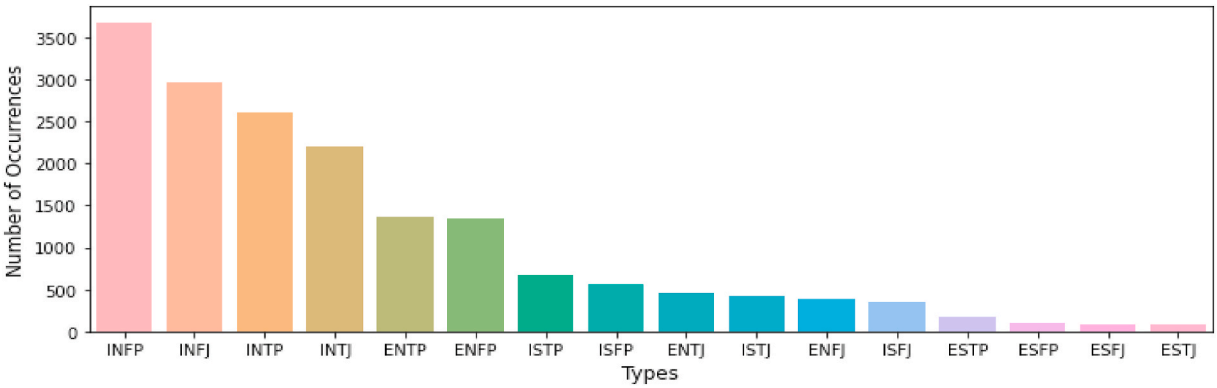

Fig. 2. Personality distribution in merged dataset.

	type	posts	IE	NS	TF	JP
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	1	1	0	1
1	ENTP	'I'm finding the lack of me in these posts ver...	0	1	1	0
2	INTP	'Good one _____ https://www.youtube.com/wat...	1	1	1	0
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	1	1	1	1
4	ENTJ	'You're fired.|||That's another silly misconce...	0	1	1	1

Fig. 3. Mapping of personalities to I-E, N-S, T-F and J-P.

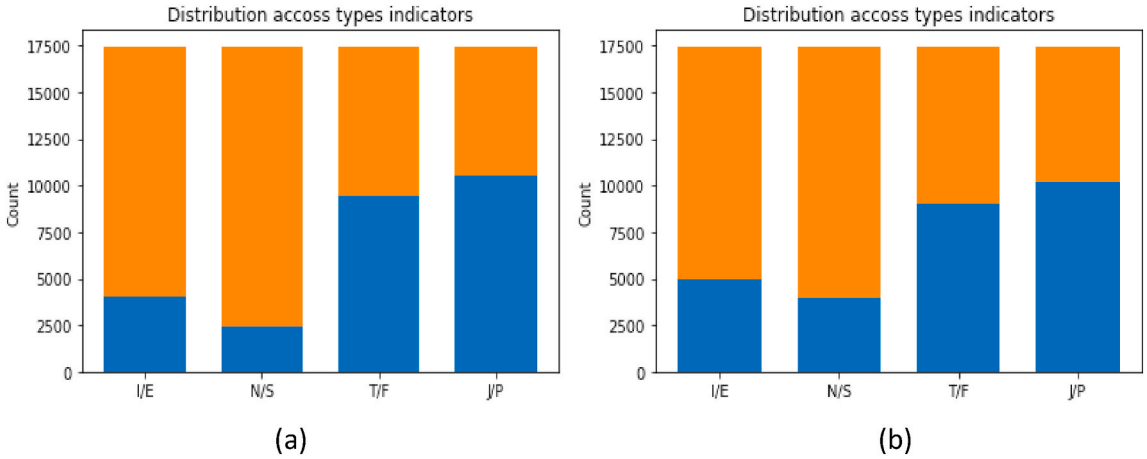

Fig. 4. Distribution of personality indicators (a) before resampling (b) after resampling.

stored in two fields: 'Type' and 'Post'. The Type field represented personality type and the Post field represented user tweets. The Post field consisted of the user's last 50 tweets stored in a comma-separated values (CSV) file and separated by the symbol |||. The initial data was collected according to the MBTI personality model to make it similar to the existing dataset. The two datasets were merged, and the distribution of the 16 personality types is shown in Fig. 2.

The 16 personality types in the datasets were mapped to eight contradictory indices. This led to four different axes: I–E, N–S, T–F and J–P. In order to do this mapping for I–E, the letter I was taken as one and E was taken as zero. Likewise, for N–S, N was taken as one and S as zero. The same was done for T–F and J–P. Fig. 3 shows the mapping of the personality types.

The distribution of eight indicators for four axes is shown in Fig. 4. It was striking to observe in Fig. 4 that the personality indicators were not evenly distributed. Therefore, the indicators were rebalanced by resampling. The less frequent indicators (i.e. I and N) were increased, whereas E and S were decreased. The same was done for T–F and J–P. The percentages of I, N, T and J before resampling were 23%, 14%, 54% and 60%, respectively. These changed to 28%, 23%, 52% and 59% after resampling, as shown in Fig. 4(a) and (b), respectively.

The axes I–E and N–S remained unbalanced. However, further resampling may lead to more repetitions of one type of index, as instances of I and N indices were only 23% and 14% in the original dataset. Axis T–F was perfectly balanced, as was J–P.

3.2. *Preprocessing and mapping of data*

Data pre-processing is a necessary step in which stop word removal, removal of personality indicators and word lemmatization were performed using Python's Natural Language Toolkit (NLTK). The portion of the dataset after pre-processing is shown in Fig. 5.

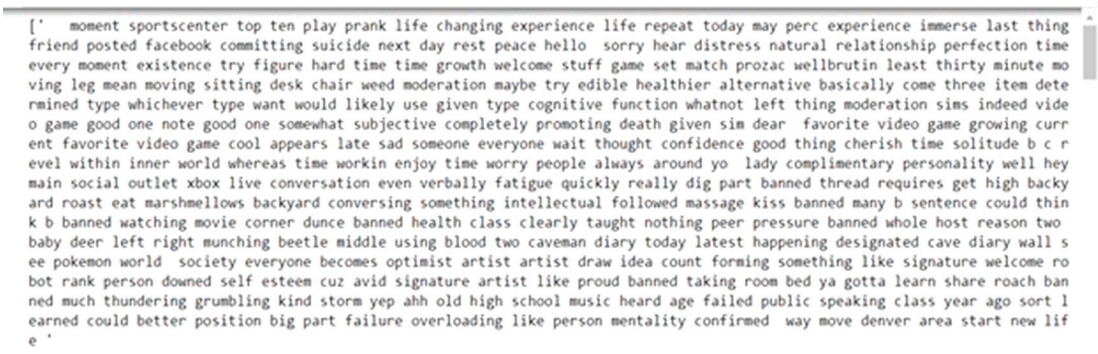

Fig. 5. Portion of dataset after preprocessing.

Table 4
Features extracted from tweets using TF-IDF.

(87, 'called'),	(234, 'family'),	(547, 'quiet'),	(769, 'worse'),
(88, 'came'),	(235, 'fan'),	(548, 'quite'),	(770, 'worst'),
(89, 'cannot'),	(236, 'far'),	(549, 'quote'),	(771, 'worth'),
(90, 'car'),	(237, 'fast'),	(550, 'random'),	(772, 'wow'),
(91, 'care'),	(239, 'favorite'),	(553, 'read'),	(773, 'write'),
(92, 'career'),	(240, 'fe'),	(554, 'reading')	(776, 'wrong'),
(93, 'case'),	(241, 'fear'),	(555, 'real'),	(777, 'wrote')
(94, 'cat'),	(242, 'feeling'),	(556, 'reality')	(783, 'yet'),
(95, 'cause'),	(243, 'fellow'),		(784, 'young'),

The dataset was then mapped to four personality terms for the BIG5 personality model using the steps given below.

1. The MBTI I-E is correlated to EXT in BIG5

   if mbti = = "i":
   EXT = 0
   elif mbti [0] = = "e":
   cEXT = 1.

2. The MBTI N-S is correlated to OPN in BIG5

   if mbti (Patel, 2020) = = "n":
   OPN = 1
   elif mbti (Patel, 2020) = = "s":
   OPN = 0.

3. The MBTI T-F is correlated to AGR in BIG5

   if mbti = = "t":
   AGR = 0
   elif mbti (Ghareb et al., 2018) = = "f":
   AGR = 1.

4. The MBTI J-P is correlated to CON in BIG5

   if mbti (Louis, 2016) = = "p":
   CON = 0
   elif mbti (Louis, 2016) = = "j":
   CON = 1.

3.3. Feature extraction

The feature extraction step helps to manage large and unstructured data collected from SM platforms (Al Marouf et al., 2020). Here, the features were extracted by term frequency and inverse document frequency (TF-IDF) (Aizawa, 2003), bag of words (BOW) (Zhang et al., 2010) and by global vectors for word representations (GloVe) embedding (Pennington et al., 2014).

3.3.1. Term frequency-Inverse document frequency(TF-IDF)

TF-IDF is a method where weights are assigned to each word present in the corpus based on its occurrence within the same and correlation in other documents. A total of 1500 features were extracted. Among these, 785 were found in the dataset and a few are shown in Table 4.

3.3.2. Bag of words (BOW)

In this method, the words present in the corpus are represented as a vector. It creates a large vector of words according to the unique words present in the dataset. Each word is assigned its frequency irrespective of its arrangement. This method of feature extraction is very easy to implement, but it creates a large and sparse vector. Moreover, it is unconcerned with the context information of words.

3.3.3. Global vectors (GloVe)

Global vectors are a dense representation of words with a consideration for context information. A global word occurrence matrix was formed for a 400 K vocabulary in 50 dimensions using a 6 B token by Stanford University in 2014 (Pennington et al., 2014). The word vectors were formed by embedding small GloVe to the collected corpus due to limitation of computing resources. The size of the corpus was defined after embedding as 48,487 rows × 51 columns. Then, the word vectors were calculated according to our corpus.

3.4. Classification

After pre-processing and feature extraction of the dataset, six machine learning algorithms were applied to classify tweets across four axes of the BIG5 personality model. To accomplish this task, the dataset was divided at an 80:20 ratio to create train and test cases.

A short description of the methods is given below.

3.4.1. Support vector machine (SVM)

This is an SML algorithm that transforms the training data into a higher dimension and builds an N-dimensional hyperplane, which splits the data into two classes using training tuples called 'support vectors. The objective of an SVM classifier is to determine the separators that best separate classes. SVM is considered an excellent classifier for text documents and is widely used in text classification where high dimensionality is the norm. However, its disadvantages include complexity and high memory requirements.

3.4.2. Decision tree (DT)

A decision tree utilizes a graph resembling a tree, or a model with decisions and their possible outcomes, along with the cost of resources and utilities to aid decision-making. It has a structure similar to a flowchart, where each internal node symbolizes a 'test' on an attribute, test outcomes are indicated by branches and the class label is

represented on leaf nodes. Rules of classification can be traced by following a path from the root to the leaf.

3.4.3. K-nearest neighbor (KNN)

The KNN model is used for learning and is driven by instance for classification. To tag text with a particular personality category, KNN considers how the majority of its neighbors voted. Various distance metrics were used for neighbor calculation, namely Euclidean distance, Makowski distance, Manhattan distance, etc. Its main advantages include its simplicity and good accuracy results if the parameters are selected sensibly. The major drawbacks of this method are the high computation and memory requirements.

3.4.4. Logistic regression (LR)

This method of SML is used to determine the probability of a categorical dependent variable. The LR model predicts binary outcome using equation (1)

$$prediction = \frac{1}{1 + e^{-(bo+bX)}} \tag{1}$$

where b_0 is the value of bias (intercept), b is the input coefficient and X is the input vector. The value of the coefficient is updated by equation (2)

$$b = b + \alpha\,(Y - \text{prediction}) \times \text{prediction} \times (1 - \text{prediction}) \times X \tag{2}$$

where $\propto$ is the rate of learning, X is the input and Y is the target variable. Lastly, the output with the highest probability is selected. The calculation of probability is stated in equation (3).

$$P(Personality|X) = \frac{1}{1 + e^{-(bo+bx)}} \tag{3}$$

3.4.5. Random forest (RF)

The random forest classifier is based on the conclusions drawn from the decision tree because of training. Let the D vector represent n decision trees for a_i as the prediction and b_i as the target variable, as shown in equation (4).

$$D = \{(a_1 b_1 \ldots \ldots ,\ a_n b_n)\} \tag{4}$$

The output of RF is acquired from the output of each decision tree $h_k(a)$ given by equation (5)

$$f(a) = f[\{h_k(a)\}] \tag{5}$$

where a is an input. Each tree casts a vote in favor of the class that is most popular at input a. The class with the maximum votes is the winner.

3.4.6. Extreme gradient boosting (XGB)

XGB is a tree-based ensemble classifier that uses the decision tree approach. Unlike other ensemble methods, it does not execute a classifier in isolation. It is an iterative approach in which the residual errors of the previous classifier are corrected by the next. This is a very powerful machine learning model, producing high accuracy in most cases.

4. Results and discussions

Three different experiments were conducted with three different feature extraction methods, and six ML classifiers were applied. All experiments were conducted on Intel Core i5-8250U CPU with 8 GB RAM except GloVe vectors embedding. GloVe feature extraction method was executed on Google Collaboratory with high RAM and GPU.

4.1. Experiment #1

The first experiment was conducted using the TF-IDF feature extraction method prior to the ML algorithms. The accuracies obtained by different classifiers are shown in Table 5 with the highest accuracy displayed in bold.

Table 5
Accuracies obtained after TF-IDF feature extraction.

	Term frequency- Inverse document frequency(TF-IDF)			
	EXT	AGR	CON	OPN
SVM	79.61	78.03	81.79	85.32
DT	89.69	85.93	93.10	92.00
KNN	92.30	96.90	96.33	95.62
LR	81.24	81.68	95.05	87.03
RF	**99.96**	98.98	**99.99**	99.88
XGB	99.84	**99.84**	99.71	**99.93**

Fig. 6 represents the accuracies obtained by experiment 1 using bar plot.

From the results of Experiment 1, it was observed that KNN, RF and XGB produce a high accuracy above 90% for all four personality types. Among them, RF and XGB are the best-performing models.

4.2. Experiment #2

The BOW feature extraction method was applied in the second experiment prior to applying the ML algorithms. The accuracies obtained by the different classifiers are shown in Table 6 with the highest accuracy displayed in bold.

Fig. 7 represents the accuracies obtained by experiment 2 using bar plot.

Table 6 and Fig. 7 show that the accuracies obtained by DT, LR, RF and XGB are above 90%. Among all, LR and XGB are the best-performing models in this experiment.

4.3. Experiment #3

Lastly, personalities were predicted using GloVe vectors in Experiment 3. The accuracies obtained by this experiment is shown in Table 7 with the highest accuracies displayed in bold.

Fig. 8 represents the accuracies obtained by experiment 3 using bar plot.

It can be observed in Table 7 and Fig. 8 that DT, RF and XGB produces accuracy above 90% and XGB perform best in this experiment. However, accuracies are generally lower than the previous two experiments. A comparison of results obtained by different feature extraction methods is presented in Fig. 9.

Nevertheless, the measurement of accuracy cannot be a suitable parameter to define its efficiency because the context information of words is more important in natural language processing (NLP). The first feature extraction method is based on term frequency, which believes that rarely occurring words are more important. The second feature extraction model also does not check for context information but collects information on the number of occurrences of a word in a document, whereas the third feature extraction model based on the global word vector embedding method, keeps the context information of words. An impact of feature extraction methods was discussed in (Ahuja et al., 2019), and classification accuracy by the TF-IDF feature extraction method was found to be higher in (Ahuja et al., 2019; Badjatiya et al., 2017).

The assessment of personalities using SM data has been a well-established research topic since 2015, shown in (Ahmad and Siddique, 2017; Kursuncu et al., 2019; Pratama & Sarno, 2015; Tadesse et al., 2018). Moreover, in the present COVID-19 situation researchers are working in sentimental analysis to assess user behavior and personalities using SM platforms (Naseem, Razzak, Khushi, Eklund, & Kim, 2021). The correlation of MBTI terms with the BIG5 personality model is being established for the Reddit dataset in (Gjurković et al., 2020). XGBoost was found to be an efficient ML classifier in (Ong et al., 2021). Vector representation of text has been found to be a suitable tool for feature

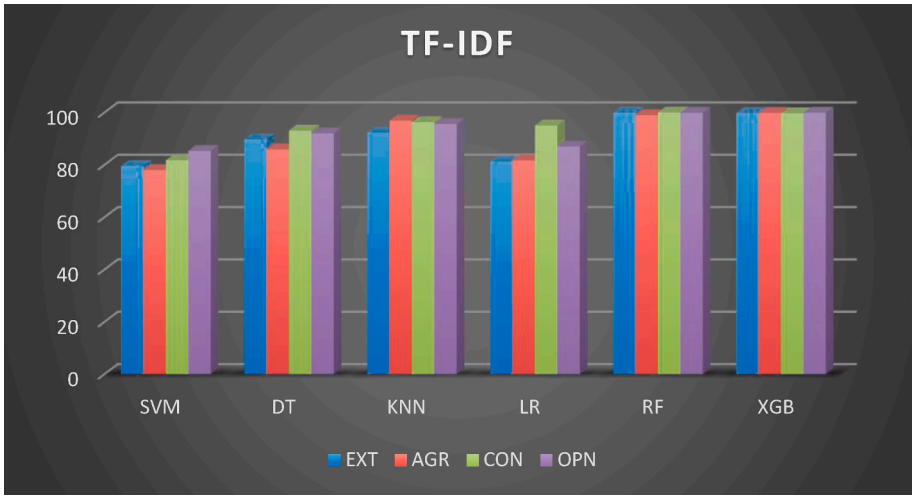

Fig. 6. Graphical representation of accuracies obtained after TF-IDF feature extraction.

Table 6
Accuracies obtained after BOW feature extraction.

	Bag of words(BOW)			
	EXT	AGR	CON	OPN
SVM	85.48	88.11	81.79	87.4
DT	95.93	94.61	93.10	96.42
KNN	79.27	71.66	69.33	86.52
LR	96.62	96.42	**95.05**	97.42
RF	94.85	95.02	91.41	96.67
XGB	**97.02**	**96.56**	94.27	**97.65**

extraction in short texts in (Mutiara et al., 2021; Škrlj et al., 2021).

Both TF-IDF and BOW produce high accuracy for classification than GloVe vector but do not keep spatial information of words. In addition, BOW is a sparse representation of words occurring in a text (Velioğlu et al., 2018). GloVe is a global word vector representation that extracts words by embedding a global word dictionary to a perspective dataset. This is found to be a better dense representation of words (Ni & Cao, 2020). A brief comparison of the present work with existing work is shown in Table 8.

5. Conclusion

Many different approaches are available for identifying personalities, such as interview-based, rating scales, self-reports, personality inventories, projective techniques and behavioural observation. This study identifies personalities from self-reported content on Twitter as users express their inner feelings on the SM platform. The dataset has been formed by merging two datasets. The first is a standard dataset available on Kaggle for MBTI personality identification and the other is collected through the Twitter API. The MBTI personality indicators were then mapped to four BIG5 personality items. Personalities were

Table 7
Accuracies obtained after GloVe feature extraction.

	Global vector(GloVe)			
	EXT	AGR	CON	OPN
SVM	76.20	73.40	59.40	85.71
DT	92.53	**91.15**	91.13	94.13
KNN	75.38	70.94	65.03	85.94
LR	76.07	73.19	62.60	85.46
RF	92.67	90.36	88.28	94.87
XGB	**92.70**	91.13	**91.15**	**95.88**

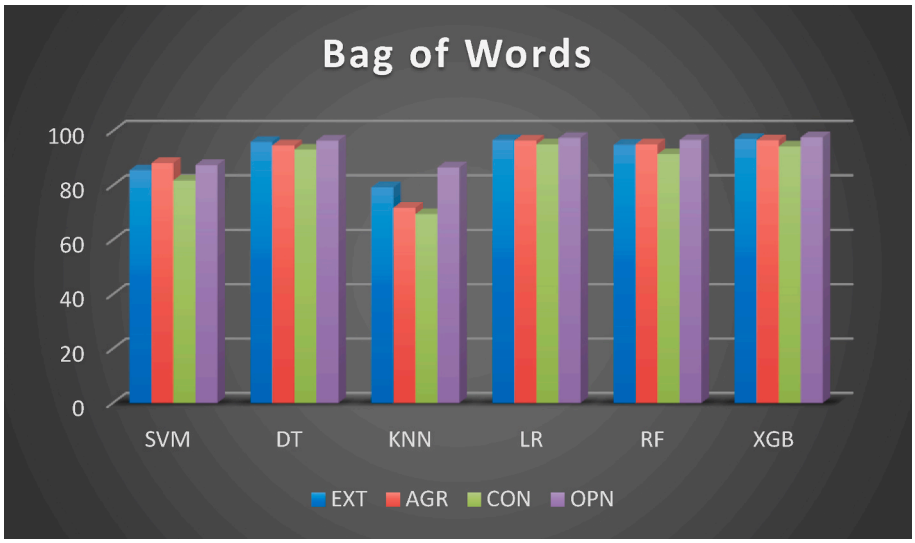

Fig. 7. Graphical representation of accuracies obtained after BOW feature extraction.

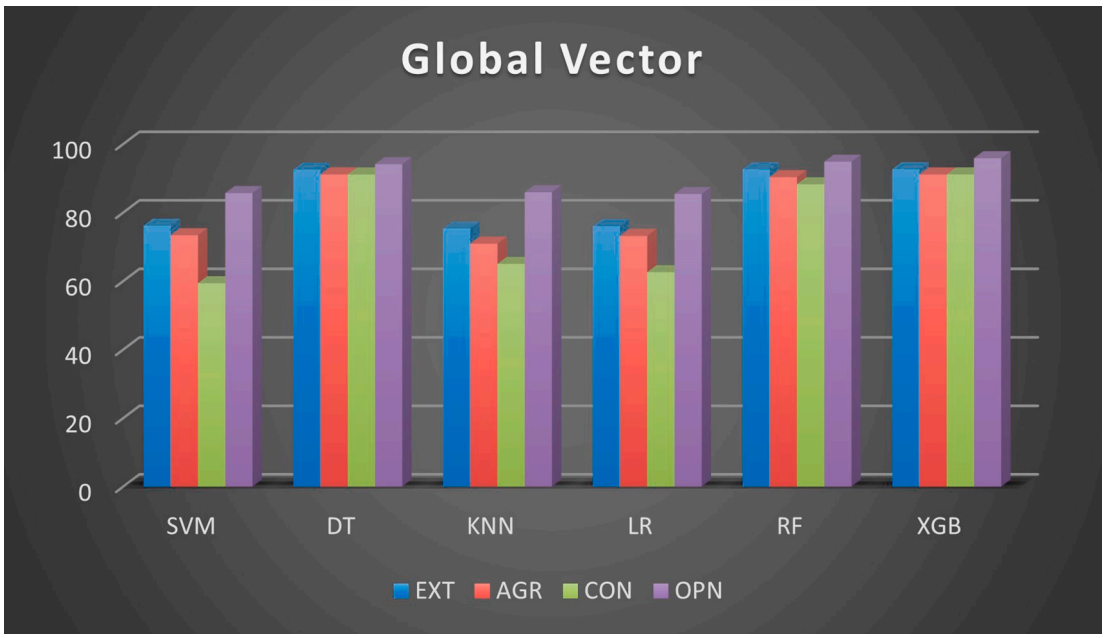

Fig. 8. Graphical representation of accuracies obtained after GloVe feature extraction.

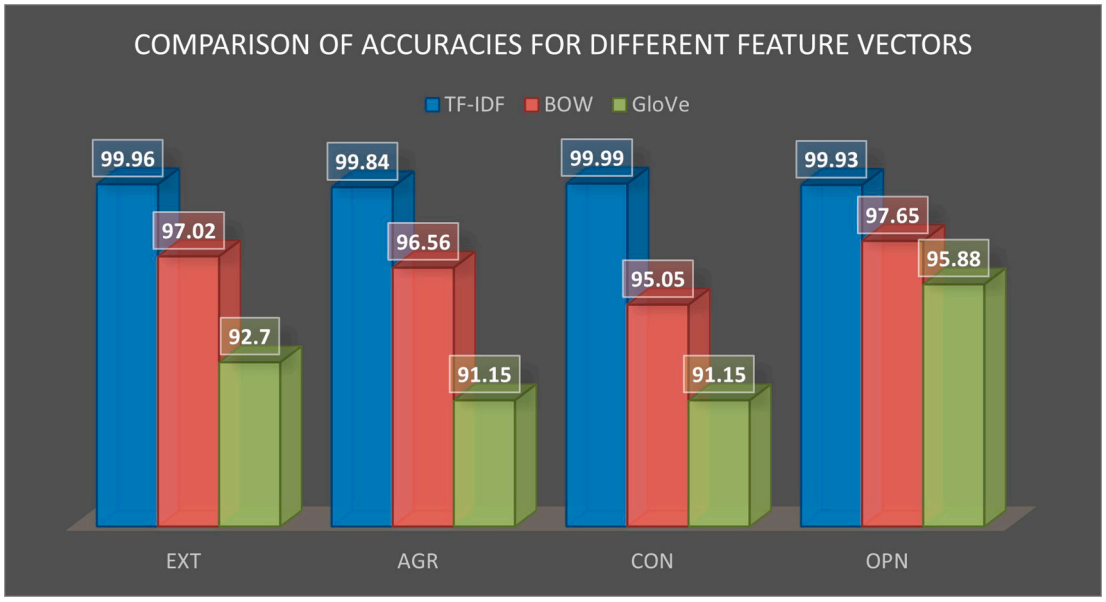

Fig. 9. Comparison of different feature extraction methods.

identified using six ML classifiers after extracting features by TF-IDF, BOW and the GloVe. The accuracy achieved by the TF-IDF feature extractor was highest. But the efficiency of a model cannot be assessed based on accuracy alone. GloVe is found to be the best feature extractor, as it is a dense representation of words with context information. The demerit of global vector representation is that it required a high computing resource. Thus, only a small GloVe vector is embedded in the present work.

The present work can be extended for data collected from different SM platforms. A combination of different features can be applied because GloVe alone produces low accuracy. Thus, TF-IDF could be modified by providing context information of words.

Funding statement

First author acknowledge research extended by the University of Delhi under IOE.

CRediT authorship contribution statement

Shruti Garg: Implementation, Writing, Compilation and Revisions. **Ashwani Garg:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 8
Comparison with existing work.

Ref, year	Dataset	Feature Extraction	Classification	Result
(Ahuja et al., 2019)	Sentiment Strength tweet dataset	TF-IDF, N Gram	SVM, LR, Naïve Bayes, RF, DT, KNN	Accuracy of TF-IDF is higher than N Gram upto 57%
(Badjatiya et al., 2017)	16 K Annotated dataset for hate speech	TF-IDF, BOW, GloVe	SVM, Gradient Boost Decision Tree (GBDT)	TF-IDF produces higher accuracy upto 81.9%
Present Work	MBTI annotated dataset 17,464 user tweets x 50 tweets per user	TF-IDF, BOW, GloVe	SVM, DT, KNN, LR, RF, XGB	TF-IDF produces highest accuracy upto 99.99%

References

Acosta, J., Lamaute, N., Luo, M., Finkelstein, E., & Andreea, C. (2017). Sentiment analysis of twitter messages using word2vec. In *Proceedings of student-faculty research day* (Vol. 7)CSIS: Pace University.

Ahmad, N., & Siddique, J. (2017). Personality assessment using Twitter tweets. *Procedia computer science, 112*, 1964–1973.

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science, 152*, 341–348.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45–65.

Al Marouf, A., Hasan, M. K., & Mahmud, H. (2020). Comparative analysis of feature selection algorithms for computational personality prediction from social -media. *IEEE Transactions on Computational Social Systems*, 1–13.

Arribas, I. P., Goodwin, G. M., Geddes, J. R., Lyons, T., & Saunders, K. E. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational Psychiatry, 8*(1), 1–7.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on world wide web companion* (pp. 759–760).

Bhavya, S., Pillai, A. S., & Guazzaroni, G. (2020). Personality identification from social media using deep learning: A review. In *Soft computing for problem solving* (pp. 523–534). Singapore: Springer.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review, 23*(2), 190–203.

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). *Pandemic populism: Facebook pages of alternative news media and the corona crisis–A computational content analysis*. arXiv preprint arXiv:2004.02566.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modelling for policy and decision making. *Policy & Internet, 7*(2), 223–242.

Celli, F., & Lepri, B. (2018). *Is Big five better than mbti? A personality computing challenge using twitter data*. In CLiC-it.

Enos, F., Benus, S., Cautin, R. L., Graciarena, M., Hirschberg, J., & Shriberg, E. (2006). Personality factors in human deception detection: Comparing human to machine performance. In *Ninth international conference on spoken language processing*.

Ezpeleta, E., Velez de Mendizabal, I., Hidalgo, J. M. G., & Zurutuza, U. (2020). Novel email spam detection method using sentiment analysis and personality recognition. *Logic Journal of IGPL, 28*(1), 83–94. https://doi.org/10.1093/jigpal/jzz073.

Furnham, A., Moutafi, J., & Crump, J. (2003). The relationship between the revised NEO-personality inventory and the Myers-Briggs type indicator. *Social Behavior and Personality: An International Journal, 31*(6), 577–584.

Ghareb, M., Karim, H., Salih, S., & Hassan, H. (2018). Social media and social relationships: A case study in kurdistan society. *Applied Computer Science, 14*(3), 31–42.

Gjurković, M., Karan, M., Vukojević, I., Bošnjak, M., & Šnajder, J. (2020). *PANDORA talks: Personality and demographics on Reddit*. arXiv preprint arXiv:2004.04460.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). November). KNN model-based approach in classification. In *OTM confederated international conferences" on the move to meaningful internet systems* (pp. 986–996). Berlin, Heidelberg: Springer.

Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 873–919). Consulting Psychologists Press.

Kaushal, V., & Patwardhan, M. (2018). Emerging trends in personality identification using online social networks—a literature survey. *ACM Transactions on Knowledge Discovery from Data, 12*(2), 1–30.

Kharde, V., & Sonawane, P. (2016). *Sentiment analysis of twitter data: A survey of techniques*. arXiv preprint arXiv:1601.06971.

Kietzmann, J., & Canhoto, A. (2013). Bittersweet! Understanding and managing electronic word of mouth. *Journal of Public Affairs, 13*(2), 146–159.

Kircaburun, K., Alhabash, S., Tosuntaş, Ş. B., & Griffiths, M. D. (2020). Uses and gratifications of problematic social media use among university students: A simultaneous examination of the big five of personality traits, social media platforms, and social media use motives. *International Journal of Mental Health and Addiction, 18*(3), 525–547.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. New York: Springer-Verlag.

Kunte, A., & Panicker, S. (2020). Personality prediction of social network users using ensemble and XGBoost. In *Progress in computing, analytics and networking* (pp. 133–140). Singapore: Springer.

Kursuncu, U., Gaur, M., Lokala, U., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2019). Predictive analysis on twitter: Techniques and applications. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 67–104). Cham: Springer.

Lee, J., & Bastos, N. (2020). Finding characteristics of users in sensory information: From activities to personality traits. *Sensors, 20*(5), 1383.

Louis, A. (2016). Natural language processing for social media. *Computational Linguistics, 42*(4), 833–836.

Mitchell, J. (2017). MBTI) Myers Briggs Personality Type Dataset available on https://www.kaggle.com/datasnaek/mbti-type.

Moreno, D. R. J., Gomez, J. C., Almanza-Ojeda, D. L., & Ibarra-Manzano, M. A. (2019). Prediction of personality traits in twitter users with latent features. In *2019 international conference on electronics, communications and computers (CONIELECOMP)* (pp. 176–181). IEEE.

Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data, 8*(1), 1–16.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, 1–13. https://doi.org/10.1109/TCSS.2021.3051189.

Ni, R., & Cao, H. (2020). July). Sentiment analysis based on GloVe and LSTM-GRU. In *2020 39th Chinese control conference (CCC)* (pp. 7492–7497). IEEE.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567.

Ong, V., Rahmanto, A. D., Williem, W., Jeremy, N. H., Suhartono, D., & Andangsari, E. W. (2021). Personality modelling of Indonesian twitter users with XGBoost based on the five factor model. *International Journal of Intelligent Engineering and Systems, 14*(2), 248–261.

Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of Computer and System Sciences, 80*(1), 57–71.

Patel, N. (2020, May 03). 17 charts that shows where social media is heading. Retrieved from https://neilpatel.com/blog/social-media-trends/.

Patrick, C., for Lumen Learning. (2020 May 10). Personality assessment. Retrieved from https://courses.lumenlearning.com/wmopen-psychology/chapter/personality-assessment/.

Pennington, J., Socher, R., & Manning, C. D. (2014). October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). EMNLP).

Pesic, D., Lecic-Tosevski, D., Kalanj, M., Vukovic, O., Mitkovic-Voncina, M., Peljto, A., & Mulder, R. (2019). Multiple faces of personality domains: Revalidating the proposed domains. *Psychiatria Danubina, 31*(2), 182–188.

Pratama, B. Y., & Sarno, R. (2015). November). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In *2015 international conference on data and software engineering (ICoDSE)* (pp. 170–174). IEEE.

Roccas, S., Sagiv, L., Schwartz, S. H., & Knafo, A. (2002). The big five personality factors and personal values. *Personality and Social Psychology Bulletin, 28*(6), 789–801.

Russell, E., & Woods, S. A. (2020). Personality differences as predictors of action-goal relationships in work-email activity. *Computers in Human Behavior, 103*, 67–79.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics, 21*(3), 660–674.

Sartonen, M., Simola, P., Lovén, L., & Timonen, J. (2020). Cyber personalities in adaptive target audiences. In *Emerging cyber threats and cognitive vulnerabilities* (pp. 175–196). Academic Press.

Sengupta, A., & Ghosh, A. (2020). Mining social network data for predictive personality modelling by employing machine learning techniques. In *Computational advancement in communication circuits and systems* (pp. 113–127). Singapore: Springer.

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., & Pollak, S. (2021). tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language, 65*, 101104.

Srivastava, D. K., & Roychoudhury, B. (2020). Words are important: A textual content-based identity resolution scheme across multiple online social networks. *Knowledge-Based Systems*, 105624.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012, December). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *2012 11th international conference on machine learning and applications* (Vol. 2, pp. 386–393). IEEE.

Sun, B., Wu, F., & Xiao, R. (2019). Automatic personality identification using students' online learning behavior. *IEEE Transactions on Learning Technologies, 13*(1), 26–37. https://doi.org/10.1109/TLT.2019.2924223.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the facebook social media platform. *IEEE Access, 6*, 61959–61969.

Thomas, S., Goel, M., & Agrawal, D. (2020). A framework for analysing financial behavior using machine learning classification of personality through handwriting analysis. *Journal of Behavioural and Experimental Finance*, 100315.

Truity. (2020 May 10). Myers & briggs' 16 personality types. Retrieved from https://www.truity.com/page/16-personality-types-myers-briggs.

Velioğlu, R., Yıldız, T., & Yıldırım, S. (2018 September). Sentiment analysis using learning approaches over emojis for Turkish tweets. In *2018 3rd international conference on computer science and engineering (UBMK)* (pp. 303–307). IEEE.

Vijay, J. A., Basha, H. A., & Nehru, J. A. (2020). A dynamic approach for detecting the fake news using random forest classifier and NLP. In *Computational methods and data engineering* (pp. 331–341). Singapore: Springer.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1–4), 43–52.