Personality Profiling: How informative are social media profiles in predicting personal information?

Anonymous ACL submission

Abstract

Personality profiling has been utilised by companies for targeted advertising, political campaigns and public health campaigns. However, the accuracy and versatility of such models remains relatively unknown. Here we explore the extent to which peoples' online digital footprints can be used to profile their Myers-Briggs personality type. We analyse and compare four models: logistic regression, naive Bayes, support vector machines (SVMs) and random forests. We discover that a SVM model achieves the best accuracy of 20.95% for predicting a complete personality type. However, logistic regression models perform only marginally worse and are significantly faster to train and perform predictions. Moreover, we develop a statistical framework for assessing the importance of different sets of features in our models. We discover some features to be more informative than others in the Intuitive/Sensory ($p = 0.032$) and Thinking/Feeling ($p = 0.019$) models. Many labelled datasets present substantial class imbalances of personal characteristics on social media, including our own. We therefore highlight the need for attentive consideration when reporting model performance on such datasets and compare a number of methods to fix class-imbalance problems.

1 Introduction

In 2023 there are over 4.59 billion social media users worldwide, constituting approximately 60% of the world's population [14]. This enables most of the world to be connected, creating an online *information environment*. The huge amounts of individual-level data provided by each user is an important aspect of social media which is unique to this type of information environment. Consequently, it is crucial for scholars to understand how this aspect of social media may impact society. There exists a need to quantify the extent to which social media can be weaponized by governments and other organisations for influence.

Every time a user enters a social media application, they leave a unique data trace – information they have posted, liked, shared, commented, even how long they have spent viewing different material on the application. We refer to this unique trace of data as a user's online digital footprint. It has been suggested that someone's online digital footprint can expose actionable information about them; including their personality profile, relationship status, political opinions and even their propensity to adopt a particular opinion or behavior [42, 26, 36, 37, 40, 38]. Cambridge Analytica was suggested to use online digital footprints to impact the result of the 2016 US election and the 2016 Brexit referendum [42]. However, the extent to which companies like Cambridge Analytica can determine this information from social media data is still questioned [26, 36, 37]. As a result, it is of interest for individuals to understand the extent of information that is attainable from their online digital footprint. This is also of key concern for governments, who seek to maintain democracies and the ethical use of such data.

We seek to determine how informative online digital footprints are in predicting Myers-Briggs personality types. This is a theoretical model comprised of four traits/dichotomies, based on Jungian theory [7, 20]. Modelling personal information about individuals using their online information has previously enabled researchers to understand the accuracy of such models. We extend this work by creating a new labelled dataset of Myers-Briggs personality types on Twitter and a statistical modelling framework which can be generally applied to any labelled characteristic of online accounts. We aim to reconsider the personality profiling and political microtargetting performed by companies like Cambridge Analytica.

First we collect a labelled dataset of accounts with self-reported Myers-Briggs personality types. We then collect a number of different features for

these accounts including social metadata features and linguistic features: LIWC [27]; VADER [18]; BERT [13]; and Botometer [33]. We then create independent logistic regression (LR), naive Bayes (NB), support vector machines (SVMs) and random forests (RF) models on each dichotomy to model the Myers-Briggs personality type of the accounts. As part of this, we consider four different weighting/sampling techniques to adjust for class imbalances. Lastly, we provide a statistical framework for analysing the importance of different features in these models. We consider the importance of features at an individual level and across groups of features for each dichotomy. Our main contributions are: (i) A labelled dataset[1] of 68,958 Twitter users along with their Myers-Briggs personality types, the largest available dataset (to our knowledge) of labelled Myers-Briggs personality types on Twitter [*reference excluded for anonymity*] ; (ii) A statistical framework to combine NLP tools and mathematical models to predict online users' personality types, which can be more broadly used to model any labelled characteristics about online accounts; (iii) A comparison of machine learning models on NLP features, and a comparison of various weighting/sampling techniques to address problems with class imbalance; (iv) Statistical methods which compare the importance of different features in NLP-based models at an individual level and across groups of features.

2 Background

Myers-Briggs [7] is the most well-known personality model, being applied in hiring processes, social dynamics, education and relationships [12, 39, 24]. The Myers-Briggs Type Indicator (MBTI) handbook illustrates a four factor model of personality where people form their 'personality type' by attaining one attribute from each of four dichotomies; Extrovert/Introvert, Intuitive/Sensory, Thinking/Feeling and Judging/Perceiving. This gives 16 different personality types where a letter from each dichotomy is taken to produce a four letter acronym, e.g., 'ENTJ' or 'ISFP'.

The model has received substantial scrutiny, particularly from psychologists who question its validity and reliability [29, 16]. Nonetheless, we utilise the Myers-Briggs model in our analysis for the following reasons: (i) Thousands of Twitter users self-report their MBTI on Twitter. This enables us to obtain a labelled dataset through appropriately querying for each of the 4 letter personality type acronyms that are unique to MBTI. (ii) The Myers-Briggs model has the largest number of self-reports on Twitter, enabling us to achieve the largest labelled personality dataset on Twitter. (iii) We aim to develop a framework for modelling personality profiles from social media data using statistical machine learning (ML) approaches. MTBI is a test case for our framework, which can be applied to other personality models (or other labelings/characteristics of individuals on social media) more generally.

Open-source labelled training data with Myers-Briggs personality types has not existed until recently. Plank and Hovy [30] modeled the MBTI of Twitter users through attaining a small dataset of 1,500 users and Gjurković and Šnajder [15] modeled the MBTI on a larger corpus of Reddit users. In 2017, Jolly [19] posted a labelled MBTI dataset on Kaggle, constituting the only known publicly available labelled dataset used for modelling the MBTI of social media users. The dataset was comprised of 8,675 users, their personality types and a section of their last 50 posts on an online forum called `personalitycafe.com`. This small online forum contains 153,000 members dedicated to discussing health, behavior, personality types and personality testing. The discussions are therefore quite different to those on other social media platforms, and likely a different demographic. Hence, this dataset is likely not generalisable to other platforms like Twitter and Facebook. It is also relatively small and imbalanced, limiting which models can be utilised on various feature sets. Class imbalance is considerable in all cases, and in one particular dataset some classes up to 28 times larger than their counterpart. Nevertheless, many papers apply machine learning models to such datasets without accounting for these class imbalances [36, 4, 21, 3, 26]. Consequently, the metrics reported often misrepresent model performance, and instead highlight the severity of class imbalances in the datasets.

3 Data Collection & Preprocessing

We discovered a number of Twitter accounts self-report their MBTI on Twitter as a regular expression. We therefore formulated two methods for querying and labelling the Myers-Briggs person-

[1] Dataset available at `https://figshare.com/s/a515f7ea420c0137f475`.

ality type of accounts. Let Ω define the set of 16 acronyms for Myers-Briggs personality types, then:

M1 Query: $\{\mathtt{x} : \mathtt{x} \in \Omega\}$. We obtained the set of users who currently self-report their personality type in their username or biography.

M2 Query: $\{(\text{I am } \mathtt{x}) \vee (\text{I am a } \mathtt{x}) \vee (\text{I am an } \mathtt{x}) : \mathtt{x} \in \Omega\}$. We obtained the set of users who have self-reported their personality type in a Tweet since Twitter's creation (March 26, 2006). Note that we only searched for self-reports in Tweets, not Retweets, Quotes and Replies – due to a number of users often not self-reporting their own MBTI when referencing MBTI acronyms in these forms of communication.

Queries were not case-sensitive.

The resulting labelled dataset comprised of 68,958 users; the dataset and more details on its collection are provided in [*reference excluded for anonymity*] . We collected 15,986 accounts by querying usernames and biographies, and 52,972 accounts from querying tweets, with misclassification rates 1.9% and 3.4% based on random samples of 1,000 accounts from each.

Next we obtained account characteristics for each user, including their: biography, most recent 100 tweets/quotes, as well as a set of Social Metadata (SM) features. The user's biography and the 100 tweets/quotes were used to generate a set of linguistic features, whereas SM features (Table 1) are directly used as numeric features in the models.

We removed duplicate users, then combined the biography and tweets into a combined text for every account. We then: 1. Normalised the text and calculated each account's dominant language. 2. Removed non-English language using the Compact Language Detect 2 (PyCLD2) library. 3. Calculated (language-dependent) Botometer scores[2]. 4. Converted text to lowercase, removed URLs, email addresses, punctuation and numbers. 5. Tokenized using the Tweet Tokenizer from the Natural Language Toolkit (NLTK) [6]. 6. Removed empty tokens and any instances of the 16 MBTI acronyms.

Next, we formulated an inclusion-exclusion criteria to determine whether a personality could be profiled from a Twitter account: we kept accounts with over 100 tweets/quotes, over 50% English language, Botometer CAP score less than 0.8, and strictly one MBTI type referenced.

We use the Botometer CAP score because we are interested in the overall bot likelihood and not the sub-category bot likelihoods. Unfortunately, there is no consistency in the literature on thresholds for binary bot classification. Rather, authors define their threshold based on a false positive rate in the context of their problem. For instance, Wojcik et al. [41] use a threshold of 0.43 for their political analysis of the twittersphere, whereas Keller and Klinger [22] use a larger threshold of 0.76 for their analysis of social bots in election campaigns. To avoid large numbers of false positive bot classifications, we chose a high threshold of 0.8.

Finally, we extracted the LIWC, BERT and VADER features from the text. The data cleaning techniques above were performed only for LIWC feature extraction, whereas the BERT and VADER features can be extracted directly from the raw text output. Thus, we calculated the LIWC features on the combined text by micro-averaging the tokens present in each LIWC category for every user. Next, we calculated the BERT features on the raw Twitter output using BERTweet [25], a pre-trained language model for English Tweets. First, we averaged the embeddings for the tokens to form a single embedding vector for each tweet/quote, then averaged the embedding vectors for the tweets/quotes to create a single 768-dimensional embedding vector for each user. We calculated the VADER features (sentiment, proportion of positive words and proportion of negative words) on the raw Twitter output for each user and include scores for both a user's biography and their tweets. We distinguish these because of contextual differences in the language; biographies often discuss oneself and tweets often discuss one's environment. We then have a total of 866 features; these are provided in Table 1.

4 Exploratory Data Analysis

We performed an exploratory data analysis (EDA) on the dataset to determine important information about our dataset, prior to any modelling. We acknowledge and discuss two forms of potential bias in our dataset: (i) only considering MBTI types on Twitter; (ii) only selecting accounts which satisfy our inclusion-exclusion criteria as well as self-report their MBTI types on Twitter. Figure 1 demonstrates these biases through bar plots showcasing the proportions of the MBTI dichotomies

[2]Further discussion: https://rapidapi.com/OSoMe/api/botometer-pro/details

Category	Features
SM	followers_count, friends_count, listed_count, favourites_count, geo_enabled, verified, statuses_count, default_profile, default_profile_image, profile_use_background_image, has_extended_profile
Botometer	cap_english, english_astroturf, english_fake_follower, english_financial, english_other, english_self_declared, english_spammer
LIWC	function, pronoun, ppron, i, we, you, shehe, they, ipron, article, prep, auxverb, adverb, conj, negate, verb, adj, compare, interrog, number, quant, affect, posemo, negemo, anx, anger, sad, social, family, friend, female, male, cogproc, insight, cause, discrep, tentat, certain, differ, percept, see, hear, feel, bio, body, health, sexual, ingest, drives, affiliation, achiev, power, reward, risk, focuspast, focuspresent, focusfuture, relativ, motion, space, time, work, leisure, home, money, relig, death, informal, swear, netspeak, assent, nonflu, filler, total_word_count
BERT	$\{e_i \; ; \; i = 1, \ldots, 768\}$
VADER	tweets_sentiment, bio_sentiment, tweets_pos_words, bio_pos_words, tweets_neg_words, bio_neg_words

Table 1: Features in our models, separated by category.

in our dataset. We compare with a study reporting MBTI proportions on Twitter [34], and with the proportion of personality types in the general population [32].

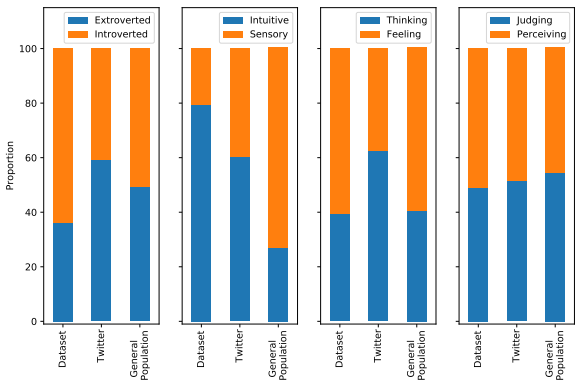


Figure 1: Proportion of accounts displaying each dichotomous trait in our dataset, on Twitter and in the general population.

A noticeable imbalance in the Intuitive/Sensory dichotomy exists across all datasets in Figure 1. There are also observable imbalances in the Extrovert/Introvert and Thinking/Feeling dichotomies. Whereas, the Judging/Perceiving dichotomy is more balanced across each dataset than the other dichotomies. The imbalances in our dataset are mostly consistent with those from www.personalitycafe.com. The higher proportion of introverts in our dataset is consistent with [23] who find that introverts tend to use social media as a primary form of communication, whereas extroverts tend to prefer communicating in-person. The larger proportion of intuitives in our dataset is consistent with Schaubhut et al. [34] who discovered that more Intuitive individuals (13%) reported being active users of Twitter than individuals with a preference for Sensing (8%). The imbalance in the Thinking/Feeling dichotomy in our dataset is opposite to what we observe in the Twitter dataset. However, Schaubhut et al. [34] found that people displaying the Feeling trait are more likely to spend their personal time browsing, interacting and sharing information on Facebook. Provided the same is true for Twitter users, our inclusion-exclusion condition requiring users to be active on Twitter (i.e. tweet/quote at least 100 times) may bias our dataset leading to more users exerting the Feelings trait.

Some authors don't assume independence between the dichotomies when modelling [4, 26], whereas most choose to model the dichotomies independently [2, 35, 5, 21, 3]. We take a data-driven approach, determining the dependency structure of the four MBTI dichotomies in our dataset using the bias-corrected version of the Cramér's V Statistic [10] (Table 2). The Cramér's V statistic is small in every case, implying that the four Myers-Briggs dichotomies are independent in our dataset, and so we model them independently.

	E/I	N/S	T/F	J/P
E/I	1.00	0.03	0.00	0.10
N/S	0.03	1.00	0.02	0.08
T/F	0.00	0.02	1.00	0.11
J/P	0.10	0.08	0.11	1.00

Table 2: Pairwise results of the bias-corrected Cramér's V Statistic between the MBTI dichotomies for our dataset.

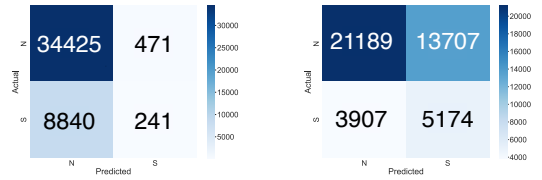
We performed a Principal Component Analysis (PCA) on the features to discover if we could significantly reduce the dimension of the feature space, and multicollinearity between the features. The first principal component explains 25.1% of the variance in the data and the first 200 principal components explain 95.4% of the variance in the data. As a result, we utilise the first 200 PCA components in our machine learning models, significantly reducing both the dimension of the feature space and the multicollinearity of the features.

5 Model Comparison

We train LR, NB, SVM and RF classifiers on each of the four dichotomies in our dataset, using 10-fold cross validation. The class imbalances we observe for some dichotomies (particularly Intu-

itive/Sensory and Extrovert/Introvert), leads us to perform four different weighting/sampling techniques prior to model fitting: (i) Weight the importance of classifying dichotomies, (ii) Upsample the minority class (with replacement), (iii) Perform the Synthetic Minority Oversampling Technique (SMOTE) on the minority class, (iv) Downsample the majority class.

Each model uses the first 200 principal components of the features in Table 1 as predictors. As an example, Figure 2 shows confusion matrices for the Intuitive/Sensory dichotomy under the standard LR model and the upsampled LR model.



(a) Standard logistic regression (b) Upsampled logistic regression

Figure 2: Confusion matrices for modelling the N/S dichotomy.

This shows that the standard LR model primarily predicts the majority class, indicating that it exploits the class imbalance to make predictions on the test sets. In comparison, the upsampled model predicts significantly more of the minority class on the test sets, resulting in more accurate predictions for the minority class. We observe similar behavior for all other models, highlighting the importance of weighting/sampling techniques to ameliorate the effect of class imbalance for prediction. However, we observe a clear trade-off between accurately predicting the majority and minority classes, with an overall reduction in accuracy due to weighting/sampling techniques. We therefore report both accuracy and Area Under the Curve (AUC) metrics for each of our models in Table 3. We report four types of accuracy depending on the number of accurately predicted dichotomies in each model. Of course, accuracy can be a misleading metric when assessing a model's performance on unbalanced data, so for comparison we report the accuracies for a random classifier and a majority class classifier. Moreover, we use an approach similar to other authors to report two types of AUC for each model [17, 11]: we macro-average and micro-average the true positive rate and false positive rate at each threshold of the ROC curve for the independent models of each dichotomy. This provides us with two ROC curves (and AUC metrics) for each model. The micro-averaged AUC aggregates the contributions of all samples in each model and weights individual predictions equally, so it is generally less sensitive to class imbalances. Table 3 compares the accuracies and AUCs of the best performing models from each method. In each case, we include the 'Standard' model and the weighted/sampling model which achieves the highest sum of micro- and macro-averaged AUC.

	Accurately Predicted Dichotomies				AUCs	
Model	4	$\geq$ 3	$\geq$ 2	$\geq$ 1	Macro	Micro
Standard LR	20.82	**60.43**	89.35	98.82	0.6688	0.6547
SMOTE LR	13.89	48.63	82.51	97.65	0.6642	**0.6620**
Standard NB	14.20	49.17	81.91	97.40	0.5784	0.5867
Upsampled NB	13.75	48.06	80.82	97.18	0.5861	0.5917
Standard SVM	**20.95**	60.25	**89.64**	**98.90**	**0.6693**	0.6518
SMOTE SVM	13.56	48.61	82.54	97.61	0.6660	0.6554
Standard RF	19.69	57.96	88.69	98.67	0.6223	0.6273
Upsampled RF	19.70	58.16	88.48	98.76	0.6305	0.6264
Random Classifier	6.250	31.25	68.75	93.75	0.5000	0.5000
Majority Class	15.31	54.54	87.20	98.28	0.5000	0.5000

Table 3: Accuracies and AUCs for best performing models. We include the 'Standard' model (with no weighting/sampling) and best performing weighted/sampling model. The 'best performing weighted/sampling model' is based on the sum of macro- and micro-averaged AUC.

Table 3 highlights the relatively small improvement in accuracy achieved by each model in comparison to the majority class classifier. It is clear that our standard SVM model is the best performing model on average. However, this model is only 5.64% more accurate at predicting a user's complete personality type compared to the majority class classifier. This is a reasonable and statistically significant improvement, but we remark based on the above discussion that the standard models are simply exploiting the class imbalances in our dataset. Moreover, we achieve similar accuracies to Plank and Hovy [30], who produced the only other Twitter dataset of labelled MBTI's (to our knowledge). In particular, we achieve better accuracies for the T/F and J/P dichotomies, and only marginally worse accuracies for E/I and N/S – further evidencing that our models perform similarly to others in the literature.

Interestingly, the standard LR model most accurately predicts at least three of four user dichotomies and is only marginally worse than SVM for all other metrics. The LR model is also significantly faster to train than the SVMs – making it the model of choice on larger datasets.

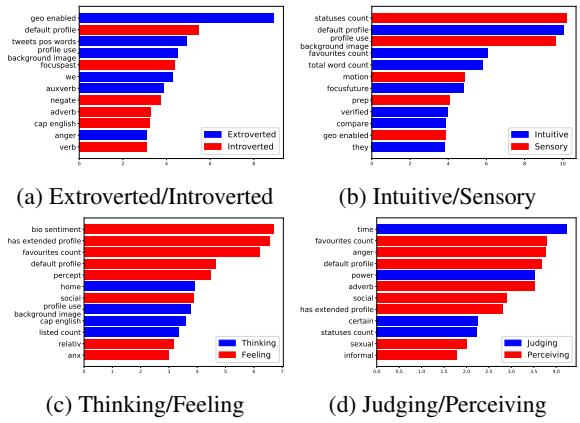
The AUC is important in discussions of model performance, especially for unbalanced datasets. This is because it equally weights the TPR and FPR, making it more robust for unbalanced datasets compared to accuracy. Most of our AUCs lie around 0.65, apart from the NB Classifiers. In particular, the best performance for the macro-averaged and micro-averaged AUCs is the standard SVM and SMOTE LR model, respectively. These AUCs are significantly larger than for both the random and majority class classifiers, indicating a clear 'signal' in our features. We therefore perform an in-depth analysis of feature importance next.

6 Feature Importance

We perform independent upsampled LR models on each of the four MBTI dichotomies because they performed well on our dataset (macro- and micro-averaged AUCs: 0.6676 and 0.6536). We choose an LR model because it is fast to train and, straightforward to interpret and perform feature selection on. Moreover, we use an upsampled model because it does not involve creating 'synthetic' data in the same way that SMOTE does – this is important for determining feature importance.

We consider the variable importance of the descriptive features in our models; these include all features except from BERT. For each dichotomy we fit the upsampled LR model and perform a stepwise feature selection to obtain a model with only significant features. In each case, we start with a null model and perform the stepwise selection algorithm on the p-values with a threshold in of 0.05 and a threshold out of 0.1. We determine the variable importance of features using the t-statistic for the parameter coefficients associated with each feature. For each dichotomy, we calculate the variable importance of each remaining feature after stepwise selection is complete, and display the absolute value of variable importance. Figure 3 displays the 12 most important features for each model. We colour the bars based on the variable's preference for each class in the dichotomy.

Pennebaker and Francis [28] suggested function words such as pronouns (pronoun), personal pronouns (ppron), 1st person singular (i), 1st person plural (we), prepositions (prep), auxiliary verbs (auxverb) and negations (negate), can describe people. Figure 3 shows the function words that are significant predictors in our models, e.g., 1st person plurals are significant in the E/I model and



(a) Extroverted/Introverted (b) Intuitive/Sensory

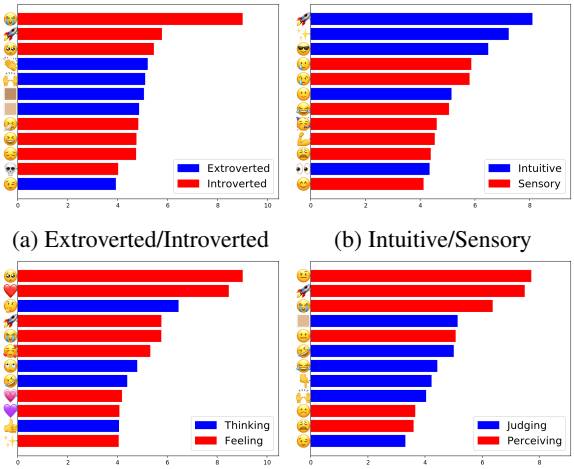
(c) Thinking/Feeling (d) Judging/Perceiving

Figure 3: Variable Importance Plots for an upsampled LR model for each dichotomy. Variables sorted by the absolute value of variable importance. Bars coloured by feature preference for each class.

prepositions are significant in the N/S model. This reinforces the importance of function words, and that techniques such as stop-word removal may remove useful information, particularly for tasks like personality prediction.

Extroverts tend to be associated with more positive language, and introverts have more focus on the past. Similarly, Chen et al. [9] suggested that extroverts display more positive emotion because they have a "dispositional tendency to experience positive emotions". Accounts with larger favourites count (i.e. the account likes more tweets) tend to be more intuitive, whereas accounts which write more statuses tend to be more sensory. Interpreting favourites as a proxy for the amount of information an account consumes, our results suggest that intuitives consume more information on Twitter, whereas sensory individuals write more. This proxy is of course not perfect, because people may consume information without liking it. Nonetheless, it is consistent with Myers-Briggs Foundation definitions, which state that intuitives pay "most attention to impressions or the meaning and patterns of the information", whereas sensors pay "attention to physical reality, what I see, hear, touch, taste, and smell" [1]. The strongest predictor for the J/P dichotomy (Figure 3d) is time; judgers are more likely to use words related to time and certainty compared to perceivers. 'End', 'until' and 'season' are examples of time-related words and 'always', 'never' are words related to certainty. This is consistent with the Myers-Briggs Foundation, which states judgers "prefer a planned or orderly way of life, like to have things settled and organized" [1].

Next we explore how emoji usage relates to a Twitter user's MBTI. On Twitter, emojis often have multiple meanings. For instance, the rainbow flag can indicate support for LGBTQ+ social movements, the wave can symbolise a "Resister" crowd of anti-Trump Twitter, and the okay symbol can be used by white supremacists, some of which covertly use the symbol to indicate their support for white nationalism [8]. Hence, emojis can indicate how these groups/movements interact with different personality types. We determine each emoji's frequency in a user's tweets and include these frequencies as predictors in upsampled LR models. Performing the same stepwise feature selection algorithm as above, we display the 12 most important predictors from the remaining models in Figure 4.



(a) Extroverted/Introverted (b) Intuitive/Sensory

(c) Thinking/Feeling (d) Judging/Perceiving

Figure 4: Variable Importance Plots for emoji counts in the upsampled LR models. Variables sorted by absolute value of variable importance. We colour bars by the feature preference for each class.

The rocket ship emoji is one the top 12 most important predictors across all models. An increase in this emoji's usage implies a higher likelihood of an account being introverted, intuitive, feelings-orientated and perceiving. The rocket ship emoji has been used by finance enthusiasts who use the emoji to denote a fast increase in a particular stock or cryptocurrency. Hence, it is possible that we are observing crypto enthusiasts to be more introverted, intuitive, feelings-orientated and perceiving. However, this emoji has other meanings like as an actual rocket ship, so we explore created word clouds of tweets containing the rocket emoji (Figure 5a), as well as the red heart emoji (Figure 5b). The rocket ship generally appears in crypto-related tweets discussing 'projects', 'great opportunities', 'developments' and 'cryptos'. However, it also appears in tweets discussing the 'moon' and 'space'. The red heart emoji mainly appears in emotive tweets discussing 'love' and 'happiness'. A number of the emojis making an account more introverted are sad/upset emojis, whereas no sad/upset emojis make an account more extroverted. This further confirms Figure 3a which suggested that extroverts prefer to display positive emotion online.

(a) Rocket Ship Emoji (b) Red Heart Emoji

Figure 5: Word clouds of tweets/quotes containing specific emojis in our dataset: rocket ship (left) and red heart (right). Note that we remove stopwords as they do not provide much context for the tweets.

Next we consider the importance of different feature groups (including the BERT features) and discuss whether different groups of features are more informative in our models. Again, we fit an upsampled LR model to all features and perform stepwise feature selection on each model. We use the same thresholds to accept and remove features. We then measure the feature group importance using the number of remaining features in each feature group after selection. For each model, Table 4 displays number of predictors (in each feature group) and proportion that remain after stepwise feature selection. This proportion can be considered a measure of the importance of each feature group, which is not biased by the number of features in each group. We introduce a statistical framework to determine whether different groups of features are more informative for our data, by performing a Chi-Squared Test on the number of features retained and excluded from each model. We test the null hypothesis that each feature group is equally informative (per feature) and include the p-values from the Chi-Square Test in the captions of Table 4.

The number of features selected depends on the type of model. For instance, 243 features are selected in the N/S model, whereas only 124 features are selected in the J/P model. Interestingly, the N/S model is also the most accurate and the J/P

Feature Type	#	Prop. Retained
SM	4	36.4%
LIWC	15	20.3%
BERT	176	22.9%
Botometer	1	14.3%
VADER	2	33.3%
Total	198	22.9%

(a) E/I ($p = 0.720$)

Feature Type	#	Prop. Retained
SM	7	63.6%
LIWC	18	24.3%
BERT	217	28.3%
Botometer	0	0.00%
VADER	1	16.7%
Total	243	28.1%

(b) N/S ($p = 0.032$)

Feature Type	#	Prop. Retained
SM	5	45.5%
LIWC	11	14.9%
BERT	124	16.1%
Botometer	1	14.3%
VADER	3	50.0%
Total	144	16.6%

(c) T/F ($p = 0.019$)

Feature Type	#	Prop. Retained
SM	4	36.4%
LIWC	8	10.8%
BERT	112	14.6%
Botometer	0	0.00%
VADER	0	0.00%
Total	124	14.3%

(d) J/P ($p = 0.120$)

Table 4: Number of features and proportion retained in each group after stepwise feature selection. p-values are from Chi-Squared Tests on the null hypothesis that each feature group is equally informative per feature.

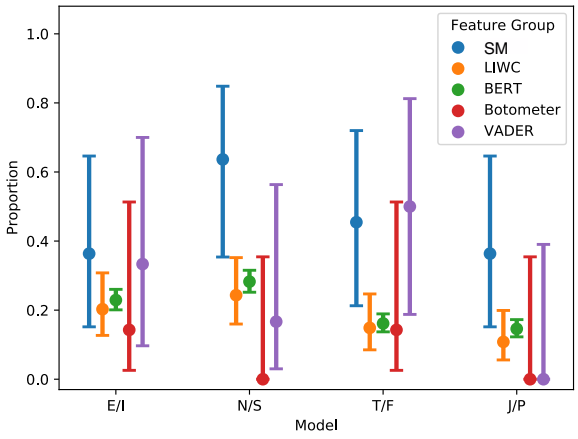


Figure 6: 95% Wilson Score Binomial CIs for the proportion of retained features in each group. We use the Wilson Score version to correct for having zero successes in some cases.

model the least accurate, implying a positive relationship between accuracy and number of features retained. This is consistent with the remark that more features are retained in a model when they are more informative about the data. Moreover, the SM features are on average the most-retained across models. Conversely, the Botometer features have worst payoff across the four models, having the smallest proportion retained on average. The most interesting comparison is between the LIWC and BERT features, which both aim to describe linguistic properties about users. In each model, the BERT features are more highly retained. However, only the results from the N/S model and the T/F model are significant at the 5% level. We therefore reject the null hypothesis that each feature group is equally as informative (per feature) for the N/S and T/F models. However, the Chi-Squared Test does not alone tell us what feature groups perform significantly better, so we perform individual confidence intervals (CIs) for the binomial proportions of accepting/rejecting features in each group using the Wilson Score interval [31]. The CIs for each feature group and model are displayed in Figure 6.

For the I/S model, the 95% CI for the SM features lies completely above those for LIWC and BERT. This indicates that SM features are more informative (per feature) than LIWC and BERT features at the 5% level for this dichotomy. Attributes about a user's account are therefore sometimes more important than the language they use when modelling personality. This is also validated by the results for the T/F model, where the 95% CI for the SM features and VADER features lie completely above the 95% CI for the BERT features. We likely observe these results because the textual features are all fairly correlated with each other. Moreover, there is no evidence to suggest that BERT features are more informative than LIWC features in determining a Myers-Briggs personality type.

7 Conclusion

This paper contributes a labelled Twitter dataset of personality types and framework to model the personality types of these users. To our knowledge, this is the largest available Twitter dataset of labelled Myers-Briggs Personality Types. Our data collection techniques avoid the long, cumbersome questionnaires used in other research. Additionally, we develop a statistical framework which combines NLP and mathematical models to model/predict users' personality type. While we applied this framework to personality types, it can model any labelled characteristics of online accounts – political opinions, psychological properties or propensity to adopt an opinion. Using this framework, we analyse and compare a number of different models. Since personality types in our dataset are unbalanced, we compare different weighting/sampling techniques to deal with class imbalances. We discover that class imbalances are common in these types of datasets, yet are often overlooked. Because of this, we demonstrate why personality prediction models appear more accurate than they are, and demonstrate why digital footprints may be less informative of personality type than models suggest.

8 Limitations

While this work provides a thorough analysis of our dataset as well as different personality models, there is certainly a need for future work in this area. Since we use a large number of features on a fairly large dataset, a deep learning model is certainly appropriate for this type of problem. Hence, it would be desirable to test the performance of our features on this dataset by utilising a suite of deep learning models. These may include models such as: Recurrent Neural Networks, Perceptron, Long Short Term Memory (LSTM) and a number of other more advanced black-box type machine learning models. These models would not give the same interpretability as the models we have used in our analysis, so they would be primarily used for their predictive capability. Class imbalances should also be acknowledged and appropriately handled in the deep learning models. Moreover, it would also be interesting to consider different methods for collecting data. One limitation of our dataset is that we only have access to the classification of the four personality dimensions, when in reality these dimensions are represented on a numerical scale. For instance, two users may be extroverted but one user may be considerably more extroverted than the other. While performing questionnaires are long and expensive, it would enable us to obtain these personality dimensions on a numerical scale, and it would reduce the mislabelling rate of the accounts. We would expect this to have a significant improvement on the performance of our models. Another obvious extension of our work is to use the OCEAN personality model instead of the Myers-Briggs model. By utilising questionnaires to obtain our data, we would have the luxury to choose which personality model to use, and so it would be possible to consider using the OCEAN model. We could then consider obtaining both personality types for each user and perform a comparison between the two personality models. This would enable us to test the reliability and accuracy of both personality models, something which has not been done by any other researchers (to our knowledge).

Acknowledgment

Excluded for anonymity.

References

[1] 2022. The Myers & Briggs Foundation - Take the MBTI® Instrument. https://www.myersbriggs.org/my-mbti-personality-type/take-the-mbti-instrument/.

[2] Firoj Alam, Evgeny A. Stepanov, and Giuseppe Riccardi. 2013. Personality Traits Recognition on Social Network - Facebook. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(2):6–9.

[3] Mohammad Hossein Amirhosseini and Hassan Kazemian. 2020. Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®. *Multimodal Technologies and Interaction*, 4(1):9.

[4] Seren Başaran and Obinna H. Ejimogu. 2021. A Neural Network Approach for Predicting Personality from Facebook Data. *SAGE Open*, 11(3):21582440211032156.

[5] Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, and Ramamoorthy Srinath. 2018. Persona Traits Identification Based on Myers-Briggs Type Indicator (MBTI) - a Text Classification Approach. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1076–1082.

[6] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, 1st edition. O'Reilly Media, Beijing ; Cambridge Mass.

[7] Melissa Block. 2018. How the Myers-Briggs Personality Test Began in a Mother's Living Room Lab. *NPR*.

[8] Conor Bronsdon. What Do Different Twitter Emojis Mean? https://conorbronsdon.com/blog/what-do-different-twitter-emojis-mean.

[9] Jiayu Chen, Lin Qiu, and Moon-Ho Ringo Ho. 2020. A Meta-Analysis of Linguistic Markers of Extraversion: Positive Emotion and Social Process Words. *Journal of Research in Personality*, 89:104035.

[10] Harald Cramér. 1946. Mathematical Methods of Statistics. In *Mathematical Methods of Statistics*, Princeton Mathematical Series ; 9. Princeton University Press, Princeton.

[11] Nunzio* Cosimo De, Luca Cindolo, Luca Sarchi, Andrea Iseppi, Mino Rizzo, Bertolo Riccardo, Andrea Minervini, Francesco Sessa, Gianluca Muto, Pierluigi Bove, Matteo Vittori, Giorgio Bozzini, Pietro Castellan, Filippo Mugavero, Daniele Panfilo, Sebastiano Saccani, Mario Falsaperla, Luigi Schips, Antonio Celia, Maida Bada, Angelo Porreca, Antonio Pastore, Al Salhi Yazan, Giampaoli Marco, Giovanni Novella, Riccardo Rizzetto, Nicoló Trabacchin, Mantica Guglielmo, Giovannalberto Pini, Riccardo Lombardo, Bernardo Rocco, Alessandro Antonelli, and

Andrea Tubaro. 2020. Using a Machine Learning Algorithm to Predict Prostate Cancer Grade. *Journal of Urology*, 203(Supplement 4):e1236–e1236.

[12] Reinout E. De Vries. 2020. The Main Dimensions of Sport Personality Traits: A Lexical Approach. *Frontiers in Psychology*, 11.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[14] Stacy Dixon. 2022. Number of Worldwide Social Network Users 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[15] Matej Gjurković and Jan Šnajder. 2018. Reddit: A Gold Mine for Personality Prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

[16] Adam Grant. Goodbye to MBTI, the Fad That Won't Die | Psychology Today. https://www.psychologytoday.com/intl/blog/give-and-take/201309/goodbye-mbti-the-fad-won-t-die.

[17] Gregor Gunčar, Matjaž Kukar, Mateja Notar, Miran Brvar, Peter Černelč, Manca Notar, and Marko Notar. 2018. An Application of Machine Learning to Haematological Diagnosis. *Scientific Reports*, 8(1):411.

[18] C.J. Hutto and Eric Gilbert. 2015. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. The AAAI Press.

[19] Mitchell Jolly. (MBTI) Myers-Briggs Personality Type Dataset. https://www.kaggle.com/datasets/datasnaek/mbti-type.

[20] C. G. Jung. 1976. *Collected Works of C.G. Jung, Volume 6: Psychological Types*, 1st edition edition. Princeton University Press, Princeton.

[21] Sedrick Scott Keh and I.-Tsun Cheng. 2019. Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-Trained Language Models.

[22] Tobias R. Keller and Ulrike Klinger. 2019. Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications. *Political Communication*, 36(1):171–189.

[23] Knowledge Leader. 2015. How Technology and Social Media Empower the Introvert. https://knowledge-leader.colliers.com/editor/how-technology-and-social-media-empower-the-introvert/.

[24] Henry W. Lane, Martha L. Maznevski, Mark E. Mendenhall, and Jeanne McNett. 2009. *The Blackwell Handbook of Global Management: A Guide to Managing Complexity*. John Wiley & Sons.

[25] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A Pre-Trained Language Model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

[26] Shankar M. Patil, Riya Singh, Paresh Patil, and Neha Pathare. 2021. Personality Prediction Using Digital Footprints. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1736–1742.

[27] James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*.

[28] James W. Pennebaker and Martha E. Francis. 1996. Cognitive, Emotional, and Language Processes in Disclosure. *Cognition and Emotion*, 10(6):601–626.

[29] David Pittenger. 1993. Measuring the MBTI ... and Coming up Short. *Journal of Career Planning and Employment*, 54.

[30] Barbara Plank and Dirk Hovy. 2015. Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal. Association for Computational Linguistics.

[31] James Reed. 2007. Better Binomial Confidence Intervals. *Journal of Modern Applied Statistical Methods*, 6(1).

[32] Michael Robinson. 1998. How rare is your personality type? https://www.careerplanner.com/MB2/TypeInPopulation.cfm.

[33] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2725–2732.

[34] Nancy Schaubhut, Amanda Weber, and Rich Thompson. 2012. Myers-Briggs Type and Social Media Report. themyersbriggs.com/contents/MBTI_and_Social_Media_Report.aspx.

[35] David Sumpter. 2018. *Outnumbered: From Facebook and Google to Fake News and Filter-Bubbles – the Algorithms That Control Our Lives*, illustrated edition edition. Bloomsbury Sigma, London.

[36] Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. *IEEE Access*, 6:61959–61969.

[37] Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio. 2017. Personality Prediction System from Facebook Users. *Procedia Computer Science*, 116:604–611.

[38] Jonathan Tuke, Andrew Nguyen, Mehwish Nasim, Drew Mellor, Asanga Wickramasinghe, Nigel Bean, and Lewis Mitchell. 2020. Pachinko prediction: A bayesian method for event prediction from social media data. *Information Processing & Management*, 57(2):102147.

[39] Bruce W. Walsh and John L. Holland. 1992. A Theory of Personality Types and Work Environments. In *Person–Environment Psychology: Models and Perspectives*, pages 35–69. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

[40] Derek Weber, Mehwish Nasim, Lucia Falzon, and Lewis Mitchell. 2020. # arsonemergency and australia's "black summer": Polarisation and misinformation on social media. In *Disinformation in Open Online Media: Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26–27, 2020, Proceedings 2*, pages 159–173. Springer.

[41] Stefan Wojcik, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. 2018. Bots in the Twittersphere.

[42] Christopher Wylie. 2020. *Mindf*ck: Inside Cambridge Analytica's Plot to Break the World*, main edition edition. Profile Trade, London.