Using SMOTE in predicting MBTI personality type based on online posts

Jeremy Dale Coronia
jeremy_coronia@dlsu.edu.ph
De La Salle University
Manila, Philippines

Jose Miguel Leyesa
jose_miguel_leyesa@dlsu.edu.ph
De La Salle University
Manila, Philippines

Luis Enrico Lopez
luis_enrico_lopez@dlsu.edu.ph
De La Salle University
Manila, Philippines

ABSTRACT

Text-based personality recognition has worked with different feature engineering techniques on various types of text, experimented with different machine learning models, and much more; however, in this experiment, we focus on addressing class imbalance to yield better accuracy results during testing. Moreover, most models in this field have used the Five-Factor Model to describe personality. In this paper, we present an opportunity to use a data augmentation technique for minority classes, SMOTE, to better represent the least represented classes and explore the extent of its effects in performing MBTI prediction based on online posts. We use SMOTE on the training set and perform feature extraction using TF-IDF. We observe the F-scores and accuracy scores to compare model performances. Results reveal that performing multi-class classification of MBTIs yield poor results. We also showed that Support Vector Machine yielded only a minimal increase in accuracy after applying SMOTE. Our analysis shows that predicting MBTI out of 16 fixed classes is significantly difficult and that applying SMOTE only yields minimal increase in performance.

KEYWORDS

natural language processing, classification, personality type, MBTI

ACM Reference Format:
Jeremy Dale Coronia, Jose Miguel Leyesa, and Luis Enrico Lopez. 2023. Using SMOTE in predicting MBTI personality type based on online posts. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The landmark works of [14, 19] have shown that writing style can be used as a reliable tool in assessing an individual's personality. The emergence of online platforms has then offered a wealth of textual data for personality research [6–8, 11, 15]. These, however, have mainly focused on the Five-Factor Model[1], a theoretical framework that measures personality using five (5) salient traits[2]. The Myers-Briggs Type Indicator (MBTI), which is dichotomic in nature, is another popular personality framework. Because MBTI is relatively unexplored in this field, we see an opportunity to explore its reliability in text classification, using blog posts.

In this paper, we investigate the use of the MBTI in predicting the personality of a user using blog posts. Specifically, we use their 50 most recent posts. Classification is performed using the 16 MBTI personality classes. We also apply Synthetic Minority Over-Sampling Technique, or SMOTE, a data augmentation technique that performs oversampling for minority classes, generating synthetic data to address significant class imbalance. We apply SMOTE only on the training set and determine how it affects model performance.

2 RELATED WORK

2.1 Personality Recognition

One of the earliest works in text-based personality recognition explored two (2) dimensions of the Big Five Model [2], showing that a given text has information that can be used to classify personality, which in this case are personality "traits" rather than "types." Further works [13, 20] focused on maximizing the information extracted from text during feature selection.

With social media and other online platforms becoming popular, personality research has shifted to online text, like with blogging sites [16, 17] and social networking sites (SNS) [7, 8, 21] because of the amount of personal data available for the assessment and prediction of an author's personality [traits]. So far, only a few have experimented with the MBTI as the psychological framework, such as with [9] who explored classification methods in personality recognition, namely, Random Forest, Extra Trees, and Gradient Boosting. Random Forest performed best in their experiment with a 45.35% accuracy, although extra trees was only close behind (45.29%), although beyond performing stratified sampling, they did not address the class imbalance. No sampling or data augmentation techniques were performed.

2.2 Myers-Briggs Type Indicator

Developed in the 1940's and widely used since then in a number of research fields and other applications, the Myers-Briggs Type Indicator (MBTI) [1] uses four (4) type indicators of individuals to generate a given personality type out of 16 classes, or sometimes also called types. The MBTI is based on Carl Jung's theory of psychological types and explored further by Katharine Briggs and Isabel Briggs Myers, both of whom created the MBTI instrument [3]. Individuals are categorized based on these type indicators. These four (4) type indicators are as follows:

[1]This is also called the Big Five.
[2]Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(1) Introversion (I) vs. Extroversion (E).
(2) Sensing (S) vs. Intuition (N).
(3) Thinking (T) vs. Feeling (F).
(4) Perception (P) vs. Judging (J).

These type indicators are then combined, forming a total of 16 unique and independent personality classes used to classify an individual.

The validity of this instrument, however, has been questioned [4, 9] because of difficulties in test-retest reliability. Individuals re-taking the MBTI questionnaire will likely not get the same result as when they first took it. Nevertheless, the MBTI instrument remains to be one of the most popular personality questionnaires used today in a variety of fields.

3 DATASET

This paper used the Myers-Briggs Personality Type (MBTI) Dataset, a publicly available dataset on Kaggle.[3] The dataset consists of a given user's 50 most recent posts, and their corresponding MBTI. It contains over 8600 rows of data consisting of: (1) the MBTI; and (2) the user's posts. The MBTI is one of the 16 unique but fixed MBTIs.

Exploratory data analysis (EDA) reveals that the MBTI classes are not equally distributed. Figure 1 shows skewness in the data. It is apparent here that the dataset is biased towards I-type (Introverted) personalities, for example, with the top 3 most represented MBTIs being I-types. On the whole, E-types (Extroverted) are under-represented.

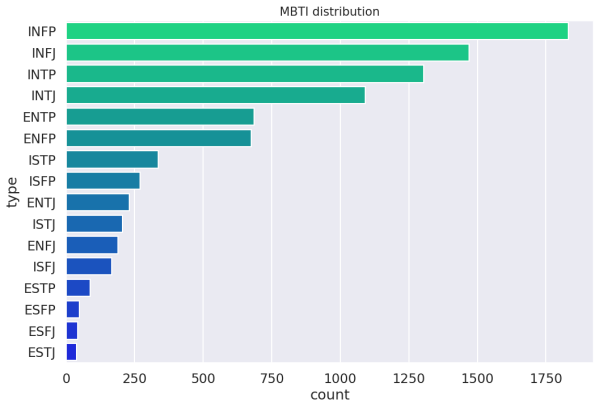

Figure 1: The distribution of recorded MBTIs. See here that the top most represented MBTIs are likely I-types while on the whole, E-types generally fall below the 500 count except for ENTP and ENFP.

To explore this further, we also plot the distribution across the type indicators to further show unequal distribution. Figure 2 shows the distribution of the type indicators. This shows significant type indicator imbalance across users.

We also explore whether or not a type indicator is independent of the other type indicators. Figure 3 shows the correlation matrix, and it can be deduced that a given type indicator does not have a strong correlation with the others.

[3]The dataset can be accessed through *this link*.

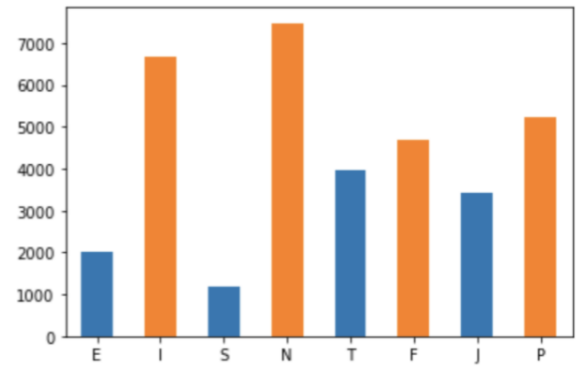

Figure 2: The distribution per MBTI type indicators. It can be observed here that I-E and N-S [type] indicators have significant inequity between them, while T-F and J-P have less of a difference.

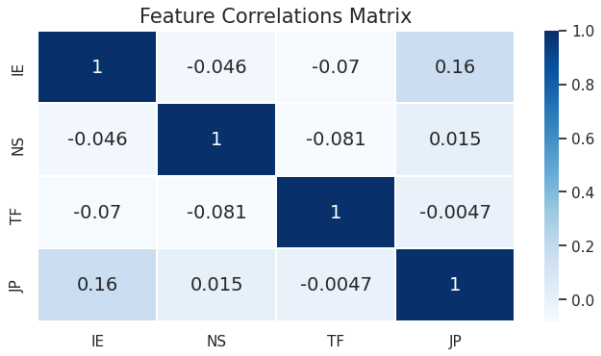

Figure 3: The heatmap of all 4 type indicators showing their correlation with the other types. The strongest positive correlation discovered is 0.16 between JP-IE, and its negative counterpart is -0.081 between TF-NS. From this, we can safely assume that type indicators are independent of each other.

The average post length (in characters) distributions were also explored. Figure 4 shows a breakdown per MBTI indicator, where it can be seen that most authors average at around the 150 character range. Note that the text has already been pre-processed prior.

3.1 Data Pre-Processing

3.1.1 Tokenization. Pre-processing is done to prepare raw text for classification. The steps are as follows: (1) Links are removed; (2) non-words are removed; (3) text is turned into lowercase; (4) mentions of MBTI classes are removed; and (5) stop words are also removed.

We perform tokenization afterwards, then perform lemmatization [12]. We also remove mentions of MBTI classes in the text because during EDA, it was discovered that about 24.27% of users self-report their MBTI type. Generated word frequency tables per

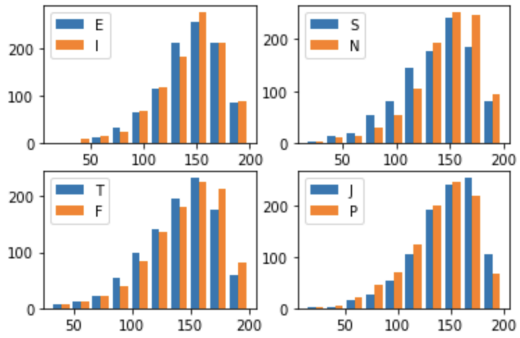

Figure 4: The distribution of post length per MBTI indicator. Post length is measured in characters, and it can be seen that on average, Thinking (T) users average shorter posts than Feeling (F) users, while Sensing (S) users average shorter posts than Intuitive (I) users.

personality type also contain MBTI classes. Removing these would validate model accuracy for unseen data.

3.1.2 Feature Extraction. We use Term Frequency Inverse Document Frequency (TF-IDF) to extract features from text. TF-IDF is divided into two (2): Term Frequency (TF) and Inverse Document Frequency (IDF). TF is used to calculate the frequency of a given term i in a document j, then divides it by the total number of words in the document, similar to a bag-of-words implementation.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

, where $n_{i,j}$ is the number of terms divided by the summation of words $\sum n_{i,j}$.

IDF, meanwhile, increases the weight of rarely used words to put it to scale with frequent terms.

$$idf(i) = log(\frac{N}{df_i})$$

,

where N is the total number of documents, divided by the number of documents where a word i occurs.

The TF-IDF is computed afterwards.

$$TF - IDF_{i,j} = tf_{i,j} * idf(i)$$

The TF-IDF is calculated by multiplying the TF with the IDF. In other words, TF-IDF calculates the relevancy of a word in a document, not just its frequency of use.

We used a threshold between 10% to 70%, which means that the calculated word occurrences within that threshold are only considered.

3.1.3 Data Augmentation. To perform data augmentation, we perform SMOTE or Synthetic Minority Over-sampling Technique [5]. SMOTE is used to synthesize new examples from the under-represented MBTI classes. This performs oversampling which can balance the distribution. SMOTE uses k-nearest neighbors to select and generate a synthetic instance.

Performing SMOTE on the data resulted in roughly 1000 instances per [MBTI] class in the training set. We did not take into consideration that there may be strong overlap between classes since we assumed that there was no strong correlation between them based on our EDA above. SMOTE was implemented using *imbalanced-learn*[4][10].

4 EXPERIMENTS

4.1 Machine Learning Algorithm

The algorithms to be used are as follows:

(1) Logistic Regression (LOG),
(2) XG Boost (XGB),
(3) Support Vector Machines (SVM), and
(4) Multinomial Naive Bayes (MNB).

The models were implemented using Scikit-Learn [18]. All settings are kept at their default configurations.

4.2 Model Training and Evaluation

Data was split into training (60%) and testing (40%). We performed a stratified split to ensure representation of each MBTI class. In observing model performance, the accuracy is used, defined as being the total number of correct MBTI predictions. The F-score is also observed to account for the class imbalance. The F-score is weighted average between the precision and recall. Precision is the ratio of true positives against all positive predictions, and recall is the measure of true positives predicted by the model.

4.3 Results and Discussion

A total of eight (8) models were created based based on the specifications discussed above. Table 1 shows the summary of results for the MBTI class models. We compare accuracy and F-score before and after applying SMOTE to determine significant differences.

Model	Non-SMOTE Accuracy	SMOTE Accuracy
XGB	25.64%	23.98%
LOG	**27.51%**	19.80%
SVM	26.92%	**27.57%**
MNB	25.11%	21.11%

Table 1: Comparison of accuracies for the MBTI class models before and after applying SMOTE.

Next, we also show F-score since the accuracy may be misleading due to the class imbalance in the test set. Table 2 shows these results before and after applying SMOTE. We also show the average precision and recall values for all models after applying SMOTE on Table 3.

4.3.1 Discussion of Results. It is apparent that the classifier struggles to perform well if it has to account for all 16 MBTIs (i.e., a multi-class problem). This can be seen in the best performance, which was 27.57% (SVM).The lowest performance is recorded at 19.80% (LOG).

Nevertheless, Logistic Regression (LOG) yielded the highest non-SMOTE accuracy with 27.51% while Multinomial Naive Bayes (MNB) performed worst with 25.11%. When SMOTE was applied,

[4]https://github.com/scikit-learn-contrib/imbalanced-learn

Model	Non-SMOTE F1	SMOTE F1
XGB	0.20	0.20
LOG	**0.23**	0.22
SVM	0.21	**0.23**
MNB	0.16	**0.23**

Table 2: Comparison of F1 values for the MBTI class models both before and after applying SMOTE.

Model	Precision	Recall	F-Score
XGB	0.20	0.24	0.20
LOG	0.25	0.20	0.22
SVM	0.23	0.28	0.23
MNB	0.28	0.21	0.23

Table 3: The weighted average precision and recall values for the MBTI class models after applying SMOTE.

LOG accuracy dipped to 19.80%, becoming the worst performing model. Our Support Vector Machine (SVM) model yielded the highest performance in this instance with 27.57%. It can be seen that implementing SMOTE for MBTI prediction overall provided only 0.67% additional accuracy for SVM; otherwise, applying SMOTE resulted in, at worst, a 7% loss of accuracy.

For F-score, the weighted average was used. Table 2 shows that the highest F-score is only 0.23 both before and after SMOTE is applied. LOG has the highest observed F-score, while both SVM and MNB both improved after SMOTE, with their F-scores also being equal. On the whole, however, the F-scores indicate low precision and recall on all models, as can be seen in Table 3. This means that our models generally performed poorly at predicting MBTI class.

5 CONCLUSION

This paper used online posts to create personality type models. We use the MBTI as our personality framework in this instance. To address class imbalance, we used SMOTE to create synthetic text data for minority classes. Experiments indicate that using the 16 MBTI classes to perform classification yields poor results, and SMOTE did not guarantee better performance. This can be observed in that only SVM resulted in better performance after applying the SMOTE technique, having the highest accuracy of 27.57%

To prevent misinterpretation of performance results due to the class imbalance, we also observe the F-score. We see that all models still yield significantly low results, with the highest F-score for both implementations only yielding 0.23, indicating poor precision and recall.

Future work can explore MBTI prediction using its type indicators (or letters) and its effect on model performance. Further testing with other data augmentation techniques available in NLP such as random insertion or back translation may also improve model performance. Experimenting with other feature extraction or pre-processing techniques may also yield better results, such as using word embeddings for input representation. Lastly, experimenting with deep learning models, both in text augmentation or in model building, can serve as another recommendation to improve prediction of MBTIs.

REFERENCES

[1] [n.d.]. https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm

[2] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W. Pennebaker. 2005. Lexical Predictors Of Personality Type. In *IN PROCEEDINGS OF THE JOINT ANNUAL MEETING OF THE INTERFACE AND THE CLASSIFICATION SOCIETY OF NORTH AMERICA*.

[3] John Beebe. 2017. *Psychological Types (Jung)*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-28099-8_625-1

[4] Gregory J. Boyle. 1995. Myers-Briggs Type Indicator (MBTI): Some Psychometric Limitations. *Australian Psychologist* 30, 1 (1995), 71–74. https://doi.org/10.1111/j.1742-9544.1995.tb01750.x

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321––357. https://doi.org/10.1613/jair.953

[6] Alastair J. Gill and Jon Oberlander. 2002. Taking Care of the Linguistic Features of Extraversion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 24. https://escholarship.org/uc/item/6n5652cx

[7] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

[8] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting Personality with Social Media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 253–262. https://doi.org/10.1145/1979742.1979614

[9] Tanay Gottigundala. 2020. *Predicting Personality Type from Writing Style*. Ph.D. Dissertation. https://digitalcommons.calpoly.edu/theses/2233/

[10] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365

[11] Jamy Li and Mark Chignell. [n.d.]. Birds of a Feather: How Personality Influences Blog Writing and Reading. *International Journal of Human-Computer Studies* 68, 9 ([n. d.]), 589–602. https://doi.org/10.1016/j.ijhcs.2010.04.001

[12] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002). http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028

[13] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Int. Res.* 30, 1 (Nov. 2007), 457–500.

[14] Matthias R Mehl and James W Pennebaker. 2003. The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology* 84, 4 (2003), 857.

[15] Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2011. Towards Discovery of Influence and Personality Traits Through Social Link Prediction. In *ICWSM-11 : Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*. 566–569.

[16] Jon Oberlander and Scott Nowson. 2006. Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics, Sydney, Australia, 627–634. https://www.aclweb.org/anthology/P06-2081

[17] Jon Oberlander and Scott Nowson. 2007. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proceedings of the 2007 International Conference on Weblogs and Social Media*. Boulder, Colorado, USA. https://www.icwsm.org/papers/2--Nowson-Oberlander.pdf

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] James W. Pennebaker and Laura A. King. 1999. Linguistic Styles: Language Use as an Indiviual Difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

[20] Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, editor="Castro Félix Howard, Newton", Alexander Gelbukh, and Miguel González. [n.d.]. Common Sense Knowledge Based Personality Recognition from Text. In *Advances in Soft Computing and Its Applications", year="2013*. Springer Berlin Heidelberg, Berlin, Heidelberg, 484–496.

[21] Edward Tighe and Charibeth Cheng. 2018. Modeling personality traits of Filipino Twitter users. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, 112–122.