Machine learning models to predict psychometric personality types to better understand customer reviews

Christopher Culley
University of Southampton
Southampton, Hampshire, UK
cc2u18@soton.ac.uk

Sam Banks
University of Southampton
Southampton, Hampshire, UK
swb1n18@soton.ac.uk

Fairoux Aldabbagh
University of Southampton
Southampton, Hampshire, UK
fa2n18@soton.ac.uk

Jak Hall
University of Southampton
Southampton, Hampshire, UK
jh16g18@soton.ac.uk

Claudia Subia
University of Southampton
Southampton, Hampshire, UK
cms2n17@soton.ac.uk

ABSTRACT

Introduction

1 INTRODUCTION

2 ETHICS

3 METHODOLOGY

3.1 Datasets

talk about the breaking INFP into four different targets.

[1] « mbti ref

4 PREPROCESSING

Biased datasets reduce the generalisability of predictive models and as such need to be corrected. The MBTi, has clear bias with some letter weightings observed for 75% of the datapoints. Although there are many options to resolve this [2], we utilise the simplest though re-sampling separate balanced datasets by target. We do this for each of the four targets in the MBTi dataset, creating four distinct train test sets with balanced targets.

In order to ready the data to build the predictive models we first apply a number of pre-processing steps in order to maximise the opportunity for meaningful patterns to be extracted. We begin by combining reviews from the same person in both the datasets into corpus from with which we apply some feature engineering listed in Table ??. We note to the reader, the sentiment analysis features, polarity and subjectivity as well as the word probability information were applied, using the NLTK library [3], to each users with the listed statistics captured over the individuals aggregated corpus.

We next turn the corpus into a bag-of-words where each instance of a word in a corpus corresponds to count in a column. We do this for the 5,000 most common words found inside the MBTi dataset ignoring stopwords. We take forward in total 5,017 features to be used inside the machine learning models.

At this point we note to the reader the quandary of scaling. Since scaling the data is a requirement for some algorithms, such as the SVM, and leads to faster solution convergence for neural networks [4] we need to apply scaling to the MBTi. In situations when it is expected that train-test splits are of the same distributions one would normally minus the train mean and divide by the train standard deviation for all data-points in both the train and test set which, in the MBTi case, presents no problem. The problem arises when we want to next apply the model learned from the MBTi to the Yelp data. Which, may follow a different distribution, this is particularly pronounced when there is more data per corpus (and thus a higher average word count). To overcome this, we normalise the Yelp data using its own mean and standard deviation which we found empirically gave a distribution of targets closer to which is expected in the general population but note this highlights a limitation in our generalisation approach – applying machine learning models from one source to another.

Features Extracted
Average word length
Max polarity
Min polarity
Average polarity
Max subjectivity
Min subjectivity
Average subjectivity
Percentage of words misspelled
Average misspelled word length
Max word probability
Average word probability
Standard deviation of word probabilities
Percentage of words that are emoticons
Percentage of letters that are punctuation
Percentage of words that are uppercase (excluding single letters)
Percentage of letters that are numerical
Percentage of words that are stop words

Table 1: The features extracted from the corpus of words defined by each user in both the MBTi and the Yelp datasets

5 MACHINE LEARNING

[5] - SVMs are good «

6 ANALYSIS

7 APPLICATION

8 DISCUSSION

[6] mbti limitiations

[7] « comparative study we could have got more ideas from

Further word will need to be done to understand the underlying distribution of the yelp and see how close it is to the mbit (kullback leiber divergence)

REFERENCES

[1] M. H. McCaulley, "The myers-briggs type indicator: A measure for individuals and groups," *Measurement and evaluation in counseling and development*, vol. 22, no. 4, pp. 181–195, 1990.

[2] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Advances in neural information processing systems*, pp. 601–608, 2007.

[3] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 31, Association for Computational Linguistics, 2004.

[4] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–48, Springer, 2012.

[5] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, pp. 137–142, Springer, 1998.

[6] G. J. Boyle, "Myers-briggs type indicator (mbti): Some psychometric limitations," *Australian Psychologist*, vol. 30, no. 1, pp. 71–74, 1995.

[7] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Icml*, vol. 97, pp. 412–420, 1997.