Using classification to determine whether personality profiles of countries affect various national indexes

Emine Yaman
International University of Sarajevo
Sarajevo, Bosnia and Herzegovina
eyaman@ius.edu.ba

Azra Musić - Kılıç
International University of Sarajevo
Sarajevo, Bosnia and Herzegovina
azramusickilic@gmail.com

Zaid Zerdo
International University of Sarajevo
Sarajevo, Bosnia and Herzegovina
zaid.zerdo@gmail.com

***Abstract*—Nations are consisted of individuals, which makes it natural that these individuals affect a nation's social structure. The effect is stronger when large groups of like-minded people are present. In this paper, we analyze do personality types, according to the MBTI theory, affect various national indexes, such as the Human Development Index (HDI), GDP per capita, Gay Happiness Index (GHI), religiosity, democracy index, etc. This is done by using four different classifiers that try to classify a country in every index provided. A few other datasets, such as the population in 1998 and surface area of a country, are used as reference points of classifiers with low classification rates. Results show that classification is possible using the national personality profile. The highest classification rates are connected with the GHI and average IQ of a country, while the lowest is scored with the reference datasets and the military budget per capita index.**

***Keywords*—data mining, classification, 16personalities, Myers-Briggs Type Indicator (MBTI), democracy index, GDP index, religiosity index, Human Development Index (HDI), GDP per capita, Gay Happiness Index (GHI), Global Peace Index (GPI)**

I. Introduction

People have always attempted to describe and categorize themselves in many ways. From the four temperaments of the Ancient civilizations – sanguine, choleric, melancholic and phlegmatic – to the latest advances in psychology, people have been restless in their pursuit of a reliable way to fit the complex human personality into a well-described model. Even though the current models can predict with a solid degree of confidence how people are likely to behave in specific situation, it is important to mention that, besides the personality, our actions are also influenced by the environment, experience, and individual goals.

The father of analytical psychology, Carl Gustav Jung, developed the theory of psychological types, which states in the 1920s was used to develop the Myers-Briggs Type Indicator (MBTI) [1]. MBTI has become as one of the most popular personality indicators used today. According to the MBTI, the four possible pairs of personality traits are:

- Introversion (I) or Extraversion (E)
- Intuition (N) or Sensing (S)
- Thinking (T) or Feeling (F)
- Judging (J) or Perceiving (P)

The theory of 16Personalities has introduced the fifth pair (Identity) of personality traits, Assertive (-A) or Turbulent (-T), the aspect that underlies all others, showing how confident people are in their abilities and decisions.

Based on the test results of more than 54 million people worldwide, the distribution of personality types by each country is determined. This data is publicly available on the 16Personalities website, in the Country Profiles subpage. On the other hand, based on the world statistics research conducted, several social indexes are assigned to each country, including Democracy Index, GDP Index, Human Development Index, and Religiosity Index. In this paper, the percentage presence of 32 different personality types is used to classify countries in different classes with respect to eight different indexes, and it examines whether the personality types distribution has effect on the democracy, human development, religiosity, and GDP status of a country.

The organization of the paper is as follows: section 2 provides a description of data and prediction models used, section 3 the experimental results obtained for each classification, and section 4 provides the overall conclusion and the scope for future research.

II. Research Method

A. Data

Dataset used is obtained from 16Personalities website, and consists of 32 attributes – each describing the presence of 32 different personality types in 124 countries worldwide.

The personality types are derived from the five personality aspects: Mind (I or E), Energy (N or S), Nature (T or F), Tactics (J or P) and Identity (T or A). It is important to understand their purpose in order to better predict the relationship with different social aspects. The Mind aspect shows how we interact with our surroundings. Introverted (I) individuals prefer solitary activities and tend to be sensitive to external stimulation such as sound, sight or smell. Extraverted (E) individuals prefer group activities and tend to be more enthusiastic than introverts. The Energy aspect determines how we see the world and process information. Observant (S) individuals are highly practical and pragmatic, focusing on what is happening or has already happened. Intuitive (N) individuals, on the other hand, are very imaginative, open-minded and curious. How we make decisions and cope with

emotions is determined by the Nature aspect. Thinking (T) individuals focus on objectivity and rationality, and tend to see efficiency as more important than cooperation. Feeling (F) individuals are sensitive and emotionally expressive, and focus on social harmony and cooperation. The forth aspect, Tactics, reflects our approach to work, planning and decision-making. Judging (J) individuals are decisive, highly organized and value planning. Prospecting (P) individuals are very good at improvising and tend to be flexible as for keeping the options open [6]. The last aspect, Identity, shows how confident we are in our abilities and decisions. Assertive (-A) individuals are self-assured and resistant to stress, while Turbulent (-T) individuals are self-conscious and sensitive to stress due to their success-driven desire to improve. The 32 personality types are shows in the Table 1 along with their roles and strategies.

TABLE I. PERSONALITY TYPES

Analysts	Confident Individualism	INTJ-A, INTP-A
	People Mastery	ENTJ-A, ENTP-A
	Constant Improvement	INTJ-T, INTP-T
	Social Engagement	ENTJ-T, ENTP-T
Diplomats	Confident Individualism	INFJ-A, INFP-A
	People Mastery	ENFJ-A, ENFP-A
	Constant Improvement	INFJ-T, INFP-T
	Social Engagement	ENFJ-T, ENFP-T
Sentinels	Confident Individualism	ISTJ-A, ISFJ-A
	People Mastery	ESTJ-A, ESFJ-A
	Constant Improvement	ISTJ-T, ISFJ-T
	Social Engagement	ESTJ-T, ESFJ-T
Explorers	Confident Individualism	ISTP-A, ISFP-A
	People Mastery	ESTP-A, ESFP-A
	Constant Improvement	ISTP-T, ISFP-T
	Social Engagement	ESTP-T, ESFP-T

According to the 16Personalities theory, personality types can be grouped into larger social types. These types include Analysts, Diplomats, Sentinels and Explorers. Each of these types serve a different role in the society. For example, Sentinels are the backbone of a society, providing stability due to their attachment to tradition, while also being the largest group of them all.

The different personality types present in each country can influence its social components that are reflected through the defined indexes. The rest of the chapter is dedicated to explaining each of the indices.

Population in 1998 [12] is dataset representing the total population of each country in 1998. The main aim of testing this dataset is to set a reference point, since marginal or no connection should be present between it and the personality profile of a country. The dataset is grouped as follows:

- Large population – larger than 20.000.000
- Medium population – between 5.000.000 and 20.000.000
- Low population – lower than 5.000.000

Surface area per country [13] is also one of the datasets to test, with the aim of setting a reference point. There should be marginal or no connection between the personality profile of a nation with the total surface area. The dataset is segmented into three classes as following:

- Large surface area – larger than 500.000 km^2
- Medium surface area – between 100.000 km^2 and 500.000 km^2
- Low surface area – lower than 100.000 km^2

Average IQ per country [14] is a dataset that contains the average score per country given on a generalized IQ test. There should be a connection between the average IQ and the personality profile, since there are papers that have connected IQ levels with MBTI groups [5]. The groups are as follows:

- High IQ – larger than 95
- Medium IQ – between 85 and 95
- Low IQ – lower than 85

GDP per capita (current $US) [15] is a dataset that reflects the ratio between purchasing power parity value of goods and services, and the average population in a given year. The categorization is made as follows:

- High GDP – larger than 20.000
- Medium GDP – between 5.000 and 20.000
- Low GDP – lower than 5.000

Human Development Index (HDI) [16] is a value assigned to countries based on life expectancy, education and fertility indicators. The index categorizes countries into three classes:

- High HDI – larger than 0.85
- Medium HDI – between 0.75 and 0.85
- Low HDI – lower than 0.75

Global Peace Index (GPI) [17] is a value given to countries according to their level of safety, security, conflict and level of militarization. The countries are grouped as follows:

- High GPI – lower than 1.6
- Medium GPI – between 1.6 and 2.1
- Low GPI – larger than 2.1

Religiosity index is a dataset developed by Gallup International [18] describing the importance of religion for each country. We segmented the dataset as follows:

- High religiosity – higher than 80%
- Medium religiosity – between 40% and 80%
- Low religiosity – lower than 40%

Military expenditure per capita [19] is an index describing the national budget spent on the military. Countries from this dataset are grouped as follows:

- High expenditure – higher than 2
- Medium expenditure – between 1 and 2
- Low expenditure – smaller than 1

Gay Happiness Index (GHI) [20] is an index indicating the level of happiness of the gay population for each country. The dataset is segmented as follows:

- High GHI – larger than 60
- Medium GHI – between 39 and 60
- Low GHI – lower than 39

B. Prediction models

1. Naïve Bayes in Weka 3.8

In machine learning, Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Bayes' theorem is given by:

$$P(y_i|x) = \frac{P(y_i)P(x|y_i)}{P(x)} \quad (1)$$

Using Bayesian probability terminology, the above equation can be interpreted as that probability of posterior is directly proportional to the multiplication of likelihood $P(x|y_i)$ and prior probability $P(y_i)$, and inversely proportional to the evidence probability $P(x)$ [7].

Weka suit offers this algorithm as well in their selection of machine learning algorithms. It can be applied to the multiclass classification problems and numeric estimator precision values are chosen based on analysis of the training data. Naive Bayes is an example of simple and fast performance algorithm that does not require high computational space.

2. Multilayer Perceptron in Weka 3.8

Weka's implementation of multilayer perceptron uses backpropagation to classify inputs into target classes. Users can control properties of algorithm including learning rate, number of neurons in hidden layer, stopping momentum, and others [8].

At the heart of backpropagation algorithm is an expression for the partial derivative ∂C/∂w of the cost function C with respect to any weight w (or bias b) in the network. The expression tells us how quickly the cost changes when we change the weights and biases. The quadratic cost function has the form

$$C = \frac{1}{2n}\sum_x \|y(x) - a^L(x)\|^2 \quad (2)$$

where: n is the total number of training examples; the sum is over individual training examples, x; y=y(x) is the corresponding desired output; L denotes the number of layers in the network; and $a^L = a^L(x)$ is the vector of activations output from the network when x is input.

After first algorithm iteration, the error in the output layer δ^L needs to be calculated using equation:

$$\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L) \quad (3)$$

After calculating error in the output layer, errors in hidden layers should be calculated as well. Equation for the error δ^l in terms of the error in the next layer δ^{l+1} is

$$\delta^l = ((w^{l+1})^T\delta^{l+1}) \cdot \sigma'(z') \quad (4)$$

After determining all the errors, gradient descent is performed in order to minimize the cost function and find optimal solution [9]. In this paper, the number of neurons in the hidden layer is set to the number produced when summation of number of attributes and classes is divided by 2 (18). The cross-validation is performed using 10-folds to test the network created.

3. Support Vector Machine in Weka 3.8

Support Vector Machine classifier in Weka, represented as SMO, implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. In that case, the coefficients in the output are based on the normalized data, not the original data. This is important for interpreting the classifier. Multi-class problems are solved using pairwise classification by Hastie and Tibshirani [10].

A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class, since in general the larger the margin the lower the generalization error of the classifier. The main aim is to maximize profit or decision function given by

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + \rho \quad (5)$$

where K is kernel of support vectors derived from the input training set of vectors and ρ is an independent term like bias is in multilayer perceptron [11].

4. RandomForest Tree in Weka 3.8

Random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset

and use averaging to improve predictive accuracy and control over-fitting [21].

Each tree is grown as follows:

1 If the number of cases in the training set is N, sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.

2 If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.

Each tree is grown to the largest extent possible. There is no pruning.

It was shown that the forest error rate depends on two things:

3 The *correlation* between any two trees in the forest. Increasing the correlation increases the forest error rate.

4 The *strength* of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.

Reducing m reduces both the correlation and the strength. Increasing it increases both [22].

III. Results

The resulting classification rates are presented in the following table, alongside the mentioned indexes:

TABLE II. Results of classification

Social component	***C1***	***C2***	***C3***	***C4***	***Avg***
Population in 1998	41%	45%	41%	51%	**44.5%**
Surface area	32%	42%	39%	40%	**38.2%**
Average IQ	59%	72%	70%	70%	**67.8%**
GDP per capita	57%	63%	65%	66%	**62.8%**
HDI	53%	57%	63%	57%	**57.8%**
GPI	51%	56%	51%	55%	**53.2%**
Importance of religion	53%	61%	65%	65%	**61.0%**
Military expenditure	37%	32%	46%	43%	**39.5%**
GHI	69%	75%	72%	75%	**72.8%**

Both the population in 1998 and surface area datasets resulted in a low classification rate, which was expected, since the personality profile of a nation is expected not to have much impact on the total population and the surface area of a country. Recent changes in the total population might be affected by the personality profile, but the total population is more affected by the past historical events and ethnicity.

Military expenditure per capita resulted in a really low classification rate. The explanation may lie in the fact that military expenditure is heavily impacted by geographical position, neighboring countries, international politics and historical events.

Both the HDI and GPI scored lower classification rates, which could be explained by the complexity of these indices, i.e. the range of factors that are taken into consideration and the various effects that affect those factors.

Importance of religion index scored modest classification rates. Various studies have been done on the connection of religiosity and MBTI, showing that personality traits, such as xSFJ, affect religiosity [3], and that ESxJ personalities showed more dogmatism [4].

GDP per capita scored solid classification rates, which might be explained by the predominance of character personalities that belong in the xSxJ-x domain (Sensing and Judging), also called the Sentinel social group, since personalities from that group are considered hardworking and responsible.

Average IQ scored quite well in the classification rate, which was expected since various papers outlined the connection between MBTI types and IQ, ex. Intuitive types scored higher than Sensing types [2].

Gay Happiness Index (GHI) scored the highest classification rate among the indexes used. This might be explained by the predominance of xNFx-x personalities, i.e. the so-called Diplomats social group. The personalities are considered to be non-traditional, pacifist, open-minded and open to change.

IV. Discussion and Conclusion

As expected, no index scored higher than 80%, which is due to the fact that a small dataset is used for classification, meaning only a few classes can be used. The other problem lies in the fact that such social constructs depend on many more factors than personality profiles.

Although the rates are average, by comparing the classification rates of the reference indices (population and area), it is possible to conclude that the personality profile of a country *does* have implications on social components and, as such, can be used to make other meaningful connections.

The next step might be to identify the key personalities that affect these and other social components. Countries can be analyzed to see in which aspects they have potential and in which they do not. Personality profiles can be used in conjunction with other data to provide better description and prediction of a country's social structure.

REFERENCES

[1] Myers, Isabel Briggs, et al. MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator. Vol. 3. Palo Alto, CA: Consulting Psychologists Press, 1998.

[2] Moutafi, Joanna, Adrian Furnham, and John Crump. "Demographic and personality predictors of intelligence: A study using the NEO personality inventory and the Myers–Briggs type indicator." European Journal of Personality 17.1 (2003): 79-94.

[3] Ross, Christopher FJ. "Jungian typology and religion: A perspective from North America." Research in the Social Scientific Study of Religion, Volume 22. Brill, 2011. 165-191.

[4] Ross, Christopher FJ, Leslie J. Francis, and Charlotte L. Craig. "Dogmatism, religion, and psychological type." Pastoral Psychology 53.5 (2005): 483-497.

[5] Kaufman, Alan S., James E. McLean, and Alan Lincoln. "The relationship of the Myers-Briggs Type Indicator (MBTI) to IQ level and the fluid and crystallized IQ discrepancy on the Kaufman Adolescent and Adult Intelligence Test (KAIT)." Assessment 3.3 (1996): 225-239.

[6] Myers, Isabel Briggs. The myers-briggs type indicator. Palo Alto, CA: Consulting Psychologists Press, 1962.

[7] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[8] "MultilayerPerceptron", *Weka.sourceforge.net*, 2016. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html. [Accessed: 17- January- 2017].

[9] M. Nielsen, *Neural Networks and Deep Learning*, "Chapter 2: How the backpropagation algorithm works", Determination Press, 2016.

[10] "SMO", *Weka.sourceforge.net*, 2016. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html. [Accessed: 16- January- 2017].

[11] L. Wang, *Support vector machines*. Berlin: Springer, 2005.

[12] United Nations Population Division: Population, total [dataset], The World Bank: IBRD [distributor]

[13] Food and Agriculture Organization: Surface area (sq. km) [dataset], The World Bank: IBRD [distributor]

[14] IQ Research: World ranking of countries by their average [dataset], IQ Research: The benchmark of IQ test [distributor]

[15] GDP per capita (current US$) [dataset], The World Bank national accounts data and OECD National Accounts data files [producer], The World Bank: IBRD [distributor]

[16] Human Development Reports: Human development index [dataset], United Nations Development Programme [distributor], 2015

[17] Global Peace Index [dataset], Institute for Economics & Peace [producer], 2016

[18] Importance of religion [dataset], Gallup [producer], 2010

[19] Military expenditure (% of GDP) [dataset], Stockholm International Peace Research Institute [producer], The World Bank: IBRD [distributor]

[20] Gay Happines Index [dataset], PlanetRomeo [producer]

[21] Breiman, Leo, and Adele Cutler. “Random Forests.” 2011. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. [Accessed: 17- January- 2017].

[22] “Random Forest”, scikit-learn.org, 2016. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: 16- January- 2017].