Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

Implementing A Naïve Bayes Classifier on Iris Data Using MATLAB, A Classification Method by Using Grid Parameters Optimization

Asmaa Ali Jasim[1], Mohammed Hasan Hadi[2]*

[1,2]Open Educational College, Ministry of Education, IRAQ

*Corresponding Author: Mohammed Hasan Hadi

DOI: https://doi.org/10.31185/wjps.516
Received 01 August 2024; Accepted 29 September 2024; Available online 30 December 2024

ABSTRACT: Data mining categorization is crucial for predicting outcomes. Naive Bayes Classification (NBC) is a prominent approach used in data mining for classification purposes. It has the ability to anticipate outcomes and is generally more efficient than other techniques of categorization. Several Naive Bayes classification methods exhibit subpar performance when used to classification and regression issues. One of the factors contributing to the success of NBC is the reliance on assumptions of independence among predictors and the initial hyper parameters. Nevertheless, this robust assumption results in a decrease in precision. This work introduces a novel approach to enhance the precision of NBC. The dataset consists of three groups, each containing 50 examples, resulting in a total of 150 cases. The dataset contains the predicted attribute for the kind of Iris plant. The findings demonstrated that the accuracy of the NBC improved to 97.81% % when Step-5 was included, compared to around 95.32 without Step-5. The optimized results are shown in Figures 2 and 3. The built-in Naive Bayes classifier (NBC) in Matlab obtained an accuracy of 78.08% using the same dataset. The suggested approach employs a grid search to enhance the accuracy of Naïve Bayes classification. The results obtained indicate that the technique utilized achieved a significant degree of accuracy in forecasting compared to the conventional built-in Matlab NBC. Thus, by assuming conditional independence, the accuracy of the NBC may be enhanced.

Keywords: Naïve Bayes Classifier, Iris Data, MATLAB, Grid Parameters, Optimization

1.INTRODUCTION

Classification is a prominent and effective technique used in data mining to forecast outcomes based on a given dataset. The Naïve Bayes Classifier is a renowned classification technique used to forecast the results of datasets. In comparison to other classifiers, the NBC demonstrates superior operational efficiency because to its exceptional predictive accuracy, straightforwardness, small memory usage, and decreased computational complexity [1].

The NBC's outstanding capacity is mostly attributed to the premise of self-reliance among the forecasters. However, this assumption and insufficiently initialized element may lead to a decrease in accuracy in the NBC. If the datasets being evaluated have qualities that strongly interact with one other, there is a greater risk of losing accuracy. Improving the precision of NBC by parameter optimization is a difficult undertaking [2].

Naive Bayes classifiers can be defined as a set of classification methods that rely on Bayes' Theorem. The algorithm is not singular, but rather a collection of algorithms that all adhere to a same principle: the independence of classification between any pair of features. Firstly, let us examine a dataset. The Naïve Bayes classifier is a straightforward and efficient classification technique that facilitates the quick building of machine learning models with fast prediction capabilities [7].

*Corresponding author: mohammed.almaawi.iq@gmail.com

The aim of this study is show that the Naïve Bayes method is used for in order solving classification issues. It is extensively used in the field of text categorization. In text classification tasks, the data is characterized by a high dimensionality, where each word represents a distinct feature in the dataset. It is used in spam filtering, emotion recognition, and rating categorization, among other applications. One of the benefits of using naïve Bayes is its efficiency. The process is rapid and the task of producing predictions is straightforward when dealing with a large amount of data [3].

This study presents a robust approach to mitigate the reduction in accuracy of the Naïve Bayes classifier resulting from insufficient initialization of the Naïve Bayes factor. The experimental trials produced results that illustrate the efficacy of the suggested method in enhancing the validity of the proposed NBC variant relative to the conventional NBC. The suggested methodology was evaluated using an IRIS dataset sourced from the UCI Machine Learning library [4].

2. LITERATURE REVIEW

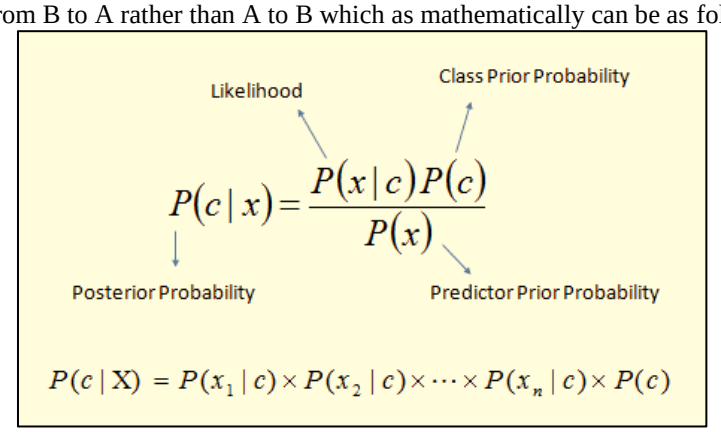
Despite its simplicity, the naïve Bayes classifier is extensively used in several practical classification and identification contexts. Additional categorization models derived from naïve Bayes have been proposed by the scientific community. The literature review examines papers that investigate the use of naïve Bayes in conjunction with diverse optimization techniques. Buratti and his colleagues developed the NACO approach, which combines a naïve Bayes classifier with an ant colony optimization algorithm [5].

The core concept of Bayes classifier can be exemplifying as follows:

We will start with a fundamental introduction to Bayes' theorem, named after Thomas Bayes from the 18th century. The Naive Bayes classifier operates on the idea of conditional probability, as articulated by Bayes' theorem. Let us examine many fundamental notions of probability that we shall use. Examine the below illustration of flipping two coins. The sample space for tossing two coins consists of the following outcomes: {HH, HT, TH, TT}. In probability calculations, we often represent probability as P [4]. The probability associated with this occurrence are as follows:

1. The likelihood of obtaining two heads is 1/4 .
2. The likelihood of obtaining at least one tail is 3/4 .
3. The conditional probability of the second coin landing heads, provided that the first coin is tails, is 1/2 .
4. The conditional chance of obtaining two heads, provided that the first coin is heads, is 1/2.

Bayes' theorem provides the conditional probability of event A, contingent upon the occurrence of event B. The first coin toss will result in B, followed by A in the subsequent toss. This may be perplexing because to the reversal of their sequence, proceeding from B to A rather than A to B which as mathematically can be as follows [6]:



This approach aims to enhance the precision of classification for datasets with a large number of dimensions. Shuangshuang Cui and his colleagues used a naïve Bayes classifier to predict the incidence of osteonecrosis in the femoral head. Sujana and colleagues presented an innovative feature selection method that integrates the naïve Bayes algorithm with the cuckoo search optimization algorithm [4].

All of the aforementioned methods encounter the issue of local optima when optimizing the hyper parameters of the naïve Bayes classifier. This work introduces a novel approach to enhance the accuracy of NBC by using grid search optimization [6]. The most famous benefits of the Naive Bayes Classifier

1. The subsequent advantages of the Naive Bayes classifier are as follows :
2. It is straightforward and uncomplicated to execute. It necessitates less training data.
3. It accommodates both continuous and discrete data. It exhibits strong scalability for the number of predictors and data points.
4. It is rapid and capable of generating real-time predictions. It exhibits insensitivity to extraneous information.

2.1 NAÏVE BAYES CLASSIFIER

Data mining relies heavily on classification. In a specific set of measurements, such as a collection of characteristic data ($x_1,x_2,\ldots,x_n$), where xi is the feature data Xi, the learning algorithm builds a classifier. The classification feature is denoted by c, and c is an instance of C in the set $c \in C \subseteq \Re m$. The goal of classification is to start groups when you have a set of observations (unsupervised learning) or when there are a lot of categories and you want to put the target into one of them (supervised learning). The classification approach used in this work is supervised learning [8].

2.1.1 NAÏVE BAYES

A classifier accurately predicts the category of a data point based on its characteristics within a certain dataset. The Naive Bayes Classifier is a supervised classification algorithm that utilizes Bayes' Theorem to predict the class of a dataset based on its properties. Naive Bayes is a probabilistic classifier that selects the class with the highest posterior probability, denoted as $\hat{c}$, out of all possible classes $c \in C$, for a given document d. In Equation 1, the use the hat notation ˆ to represent "our estimation of the accurate category" [9].

$$c^{\wedge} = \arg\max_{c \in C} P(c/d) \qquad (1)$$

The concept of Bayesian inference has been recognized since the research conducted by Bayes in 1763. It was first used in the field of text categorization by Mosteller and Wallace in 1964. The underlying principle of Bayesian classification is to use Bayes' rule in order to convert Equation (1) into alternative probabilities that possess advantageous characteristics. Bayes' theorem is expressed in Equation (2). It provides a method to decompose any conditional probability *P(x|y)* into three distinct probabilities:

$$P(x/y) = \frac{P(y/x)p(x)}{p(y)} \qquad (2)$$

Substituting Equation .2 into Equation 1 yields Equation 3:

$$c^{\wedge} = \arg\max_{c \in C} P(c/d) = \arg\max_{c \in C} \frac{P(y/x)p(x)}{p(y)} \qquad (3)$$

To simplify Eq. (3), may readily omit the denominator *P(d)*. This is feasible since this will be calculating for every conceivable category. However, the probability *P(d)* remains constant for each class. It has been consistently enquire about the most probable class for the same document d, which must possess the same probability *P(d)*. Therefore, we may choose the class that maximizes this more concise formula:

$$c^{\wedge} = \arg\max_{c \in C} P(c/d) = \arg\max_{c \in C} P(\mathrm{d}/\mathrm{c})\,p(\mathrm{c}) \qquad (4)$$

Therefore, the calculate the most likely class $\hat{c}$ for a given document d by selecting the class with the largest product of two probabilities. The text describes the prior probability of the class *P(c)* and the likelihood of the document *P(d|c)*. where $P(\mathrm{d}/\mathrm{c})$ represent the likelihood factor and $p(\mathrm{c})$ represent the prior factor. To maintain generality, it may may express the factor *d* as a collection of features (*f1, f2,....., fn*):

$$c^{\wedge} = \arg\max_{c \in C} P(f1, f2, \ldots\ldots\ldots, fn/c) \qquad (6)$$

The equation (6) represents the likelihood function. Regrettably, Equation 6 remains excessively challenging to calculate directly. Without certain simplifying assumptions, the task of calculating the likelihood for each conceivable combination of data, such as every potential arrangement of words and locations, would need an immense number of parameters and training sets of unattainable magnitude. Naive Bayes classifiers therefore rely on two simplifying assumptions. The first assumption is the bag of words assumption, which disregards the location of words in a text [10].

2.1.2 ASSUMPTION AND CATEGORIZATION

According to previous equations the assumption, the word "book" would have the same impact on categorization regardless of whether it appears as the first, 20th, or final word in the document. Therefore, it make the assumption that the characteristics (f*1, f2,……., fn*) only represent the identification of the words and do not include information about their location. The second assumption, known as the naïve Bayes assumption, states that the probabilities *P(fi |c)* are independent given the class *c*. This allows us to multiply these probabilities in a straightforward manner [11].

$$c^{\wedge} = \arg\max_{c \in C} P(f1, f2, \ldots\ldots\ldots, fn/c) \qquad (6)$$

The following equation is the final equation for the class that a naive Bayes classifier has selected:

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{f \in F} P(f / c) \quad (7)$$

Taking into account word locations is necessary in order to apply the naive Bayes classifier to text. This may be accomplished by simply traversing an index across each and every word position in the document: The places of all the words in the test paper are numbered [12].

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{i \in positions} P(w_i / c) \quad (8)$$

Calculations for Naive Bayes, as well as calculations for language modelling, are performed in log space in order to prevent underflow and to boost the speed of the process. As a result, equation 8 is often represented as:

$$c_{NB} = \arg\max_{c \in C} P(c) + \sum_{i \in positions} \log P(w_i / c) \quad (9)$$

In order to calculate the predicted class as a linear function of the input features, Equation (9), which takes into account features in log space, is used. Classifiers that produce a classification conclusion by using a linear combination of the inputs are referred to as linear classifiers. Examples of linear classifiers are naive Bayes and logistic regression [7].

2.3 HISTORICAL CONTEXT OF IRIS DATASET

The Iris dataset has great historical importance as it serves as a fundamental dataset in the fields of statistical analysis and machine learning. Ronald Fisher's analysis of the dataset laid the foundation for the development of several categorization methods that continue to be used in contemporary times. The dataset has shown to be enduring and remains a standard for evaluating new machine learning models [13].

The Iris dataset is essential in machine learning as a standardized baseline for evaluating classification systems. It is often used to showcase the efficacy of algorithms in resolving classification issues. Researchers use it to assess and contrast the efficacy of various algorithms and gauge their accuracy, precision, and recall [14]. There are several factors contributing to the widespread use of this dataset:

1 -SIMPLICITY: The Iris dataset is very significant in the field of machine learning because of its straightforwardness. Novices find it quite beneficial for comprehending essential machine learning principles such as data preparation, model construction, and evaluation. The fundamental framework of this structure comprises numerical characteristics like as sepal and petal measurements, making it clearly understandable.

2 -VERSATILITY: The Iris dataset exhibits clear distinctions across its classes - Iris setosa, Iris versicolour, and Iris virginica - despite its fundamental characteristics. This capability enables the use of many classification techniques, including logistic regression, support vector machines, decision trees, and others [15].

3 -BENCHMARKING: The Iris dataset is very beneficial as a benchmark for comparing the performance of different machine learning techniques. Scientists use this dataset to assess the effectiveness and precision of various techniques in a standardized environment, assisting in the determination of the best appropriate algorithm for certain jobs.

4 -INSTRUCTIONAL TOOL: The Iris dataset is included into the regular machine learning curriculum and is very beneficial as an instructional tool. It allows students to participate in practical learning experiences, where they may experiment with algorithms and approaches in a simple setting. This helps them better understand how theoretical ideas can be used in real-life situations.

5 -UNDERSTANDING FEATURE IMPORTANCE: The Iris dataset helps us grasp the importance of features in classification problems by providing a small number of characteristics. By directly observing, learners may get a deep understanding of how different characteristics affect the ability of a model to make accurate predictions. This allows them to comprehend important ideas linked to selecting the most relevant features and reducing the complexity of the data [16].

6- STANDARDIZATION: The Iris dataset is widely acknowledged as a standardized and internationally accepted dataset in the field of machine learning. This promotes convenient agreement among researchers when evaluating the effectiveness of various algorithms, guaranteeing a shared comprehension of the anticipated results for this dataset.

The dataset of this research consists of three groups, each containing 50 examples, resulting in a total of 150 cases. The dataset contains the predicted attribute for the kind of Iris plant

3. APPLICATIONS OF IRIS DATASET

Researchers along with data scientists use the Iris dataset for a multitude of purposes, such as:

1 -CLASSIFICATION: The Iris dataset is often used for jobs involving classification. The objective is to determine the species (classes) of an iris flower based on its four traits. The dataset may be used to train machine learning algorithms, such as decision trees, support vector machines, k-nearest neighbors, and neural networks, to accurately categories iris flowers according to their species [17].

2 -DIMENSIONALITY REDUCTION: Due to the limited number of features in the Iris dataset, it is not considered to be high-dimensional. Nevertheless, it continues to be used for the purpose of demonstrating dimensionality reduction methods, such as principal component analysis (PCA). Principal Component Analysis (PCA) may be used to decrease the dimensionality of the information while retaining a significant portion of its variance, hence facilitating visualization or analysis.

3 -EXPLORATORY DATA ANALYSIS: Conducting Exploratory Data Analysis (EDA) involves examining the distribution of features, exploring correlations between variables, and identifying any outliers present in the dataset.

4- FEATURE SELECTION: Feature selection involves identifying the key elements that significantly impact classification accuracy. The Iris dataset is often used to showcase and evaluate various feature selection strategies. The purpose of these strategies is to determine the most significant characteristics (namely, sepal length, sepal width, petal length, and petal width) that have the greatest impact on the prediction accuracy of a model [16].

4.METHODOLOGY

Initially, it is essential to recognize that the joint distributions of dependent variables might become very intricate. Managing joint distributions of many variables is among the most challenging issues in statistics and probability. Mathematical problems often simplify when we assume the independence of variables. Conversely, supposing independence entails disregarding any interactions among the effects denoted by the variables. In designing probability models, there is often a trade-off between simplicity (e.g., assuming independence) and precision (endeavoring to describe all interactions correctly).

4.1 The Data Set Studded in This Research

Among the datasets that are used most often in the area of machine learning and statistics, the Iris dataset is among the most well-known and extensive. This article will provide an in-depth examination of the Iris dataset, as well as information on its many applications and uses. The approach used in this work utilized the IRIS dataset obtained from the EITCA Machine Learning Repository. The dataset is categorized as multivariate since it contains statistical data about the Iris plant species, characterized by four specific attributes: petal width, petal length, sepal length, and values. The dataset has three groups, each with 50 occurrences, totaling 150 cases. The information includes the anticipated characteristic for the kind of Iris plant.

Table (4,1): IRIS Dataset

Row	ID	Label	al	a2	a3	a4
1	Id-1	Iris-setosa	5.100	3. 500	1. 400	0. 200
2	Id-2	Iris-setosa	4.900	3.000	1. 400	0 .200
3	Id-3	Iris-setosa	4.700	3.200	1.300	0.200
4	Id-4	Iris-setosa	4.600	3.100	1 500	0.200
5	Id-5	Iris-setosa	5.000	3.000	1 400	0 200
6	Id-6	Ids-setosa	5.400	3.000	1 700	0 400
7	Id-7	Ids-setosa	4.600	3.400	1 400	0 300
8	Id-8	Iris-setosa	5.000	3.400	1.500	0.200
9	Id-9	Iris-setosa	4.400	2.900	1 400	0.200
10	Id-10	Ids-setosa	4.900	3.100	1 500	0.100

4.2 PROCESS OF IMPLEMENTATION

To retrieve the Iris dataset, one may use the 'load_iris' method from the 'sklearn.datasets' module. This function enables the loading of the Iris dataset. The load_iris method and assign the resulting dataset object to the variable 'iris'. The object encompasses the whole dataset, including both the characteristics and the target variable. It can follow the following steps:

Step 1: "The Iris dataset in CSV format is used as the input".

Step 2: "Partition the data into separate training and testing datasets. The dataset in this research was partitioned into 70% for training and 30% for testing".

Step 3: "Separate the training dataset according to the class values, namely 1, 2, and 3".

Step 4: "Calculate the standard variation and mean values for each data case using the class values".

Step 5: "Select the laplace_correction parameter for the naïve Bayes method as the input for the grid search optimization process".

Step 6: "Utilise the optimum value of the laplace_correction parameter as a starting value for the classification process using naïve Bayes".

Step 7: "Apply the model and provide forecasts".

Step 8: "Calculate the prediction accuracy by comparing the class data of the test dataset. This accuracy is assessed by calculating the ratio on a scale of 0 to 100%". The results described in the table (4,1).

5. EXPERIMENT RESULTS

Iris setosa, Iris virginica, as well as Iris versicolour are the three unique species of iris that are represented in the Iris Dataset, which has fifty samples of each species. The length and width of the sepals and petals are the four criteria that are used to describe each sample. This dataset is often used for the purposes of data mining, classification, and clustering examples, in addition to being utilised for the testing of algorithms.

Performance Vector:
accuracy: 95.33% +/- 4.27% (mikro: 95.33%)
Confusion Matrix:

True: Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa: 50	50	0
Iris-versicolor:	0	47	4
Iris-virginica: 0	3	46

Figure (5,1): The precision achieved is 95.33% while excluding Step-5.

Performance Vector:
accuracy: 97.81%
Confusion Matrix:

True: Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa: 11	0	0
Iris-versicolor:	0	21	0
Iris-virginica: 0	1	12

Figure (5,2): The precision achieved is 97.81% while excluding Step-5.

The proposed model, as described in Section IV, was applied to the Iris dataset both with and without Step-5. After each iteration, the outcomes were assessed by measuring the accuracy of the Naive Bayes Classifier (NBC). The findings demonstrated that the accuracy of the NBC improved to 97.81% % when Step-5 was included, compared to around 95.32 without Step-5. The optimised results are shown in Figures 2 and 3. The built-in Naive Bayes classifier (NBC) in Matlab obtained an accuracy of 78.08% using the same dataset.

Table 1 displays the comprehensive examination of the proposed approach in terms of its accuracy compared to other techniques. The findings demonstrated that the proposed method effectively mitigated the decrease in accuracy in the NBC by assuming conditional independence. Therefore, the approach described in this research may improve the implementation of the NBC. The results of the suggested technique demonstrate the effectiveness of using grid search optimisation to determine the optimal hyperparameters for the naïve Bayes classifier. The grid search method divides all parameters into a defined range, computes the parameter values for every point on the grid, and identifies the ideal parameters based on accuracy.

Table (5,1): Performance Comparative Analysis

Classifier	Performance on 150 Data Instances		
	"Training (67%)"	"Test (33%)"	"Accuracy %"
NBC with a conditional independence and the fifth phase of execution.	100	50	97.81%
NBC with conditional independence wihout the fifth phase of execution.	100	50	95.32
The built-in NBC algorithm in Matlab	100	50	78.08

6. CONCLUSION

The initial value is the primary determinant of accuracy degradation in the NBC. Nevertheless, this assumption simplifies the process of estimating probabilities. The research used a separation strategy that enhanced the precision of the classifier by the implementation of grid search optimisation. The findings obtained demonstrate that the used approach attained a notable level of accuracy in predicting compared to the traditional built-in Matlab NBC. Therefore, by assuming conditional independence, the accuracy of the NBC may be improved.

REFERENCES

[1] N. G. and A. Sharma, "Performance analysis of variants of naïve Bayes classifier to identify classes of flowers using iris recognition system," in 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), 2023.

[2] A. S. Fitrani and H. Prasetyo, "Sentiment Analysis before Presidential Election 2024 Using Naïve Bayes Classifier Based on Public Opinion in Twitter," 2023.

[3] C. H. Pratama and Y. Findawati, "Hate Speech and Emotions Classification in Indonesian Language Texts on Twitter Using Naïve Bayes Classifier," 2024.

[4] A. Amitha and M. Meleet, "A system for recommendation of medication using Gaussian naïve Bayes classifier," International Journal of Innovative Research in Computer Science & Technology, vol. 7, no. 3, pp. 100–103, 2019.

[5] K. J. D'souza and Z. Ansari, "Big Data Science in building medical data classifier using naïve Bayes model," in 2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2018.

[6] Venkatesh and K. V. Ranjitha, "Classification and Optimization Scheme for text data using machine learning naïve Bayes classifier," in 2018 IEEE World Symposium on Communication Engineering (WSCE), 2018.

[7] I. Simatupang, D. S. Pamungkas, and S. K. Risandriya, "Naïve bayes classifier for hand gestures recognition," in Proceedings of the 3rd International Conference on Applied Engineering, 2020.

[8] E. Azeraf, E. Monfrini, and W. Pieczynski, "Improving usual naive Bayes classifier performances with neural naive Bayes based models," in Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, 2022.

[9] J. Luengo and R. Rumi, "Naive Bayes classifier with mixtures of polynomials," in Proceedings of the International Conference on Pattern Recognition Applications and Methods, 2015.

[10] R. M. Moraes and L. S. Machado, "A fuzzy Poisson naive Bayes classifier for epidemiological purposes," in Proceedings of the 7th International Joint Conference on Computational Intelligence, 2015.

[11] D. Berrar, "Bayes' theorem and naive bayes classifier," in Encyclopedia of Bioinformatics and Computational Biology, 2019, pp. 403–412.

[12] P. Subarkah, W. R. Damayanti, and R. A. Permana, "Comparison of correlated algorithm accuracy naive Bayes classifier and naive Bayes classifier for classification of heart failure," ILKOM Jurnal Ilmiah, vol. 14, no. 2, pp. 120–125, 2022.

[13] A. Madrid, "Context and historical perspective of the conflict in Honduras," PsycEXTRA Dataset, 2013.

[14] J. Arditti, "Historical and contemporary context of maternal and child rights during incarceration," PsycEXTRA Dataset, 2014.

[15] V. Sewell, "Untapped potential in a secondary context," Iris Journal of Scholarship, vol. 1, pp. 108–117, 2019.

[16] L. Logan, "Addressing vicarious trauma in the context of historical trauma: A relational world view model," PsycEXTRA Dataset, 2011.

[17] J. Kuriansky, "Magic in the City of Novosibirsk: Historical and cultural context to the international expert symposium," PsycEXTRA Dataset, 2014.