Multimodal Technologies and Interaction

MDPI

Article

A Machine Learning Approach to Identify the Preferred Representational System of a Person

Mohammad Hossein Amirhosseini * and Julie Wall

Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London, London E16 2RD, UK
* Correspondence: m.h.amirhosseini@uel.ac.uk

Abstract: Whenever people think about something or engage in activities, internal mental processes will be engaged. These processes consist of sensory representations, such as visual, auditory, and kinesthetic, which are constantly being used, and they can have an impact on a person's performance. Each person has a preferred representational system they use most when speaking, learning, or communicating, and identifying it can explain a large part of their exhibited behaviours and characteristics. This paper proposes a machine learning-based automated approach to identify the preferred representational system of a person that is used unconsciously. A novel methodology has been used to create a specific labelled conversational dataset, four different machine learning models (support vector machine, logistic regression, random forest, and k-nearest neighbour) have been implemented, and the performance of these models has been evaluated and compared. The results show that the support vector machine model has the best performance for identifying a person's preferred representational system, as it has a better mean accuracy score compared to the other approaches after the performance of 10-fold cross-validation. The automated model proposed here can assist Neuro Linguistic Programming practitioners and psychologists to have a better understanding of their clients' behavioural patterns and the relevant cognitive processes. It can also be used by people and organisations in order to achieve their goals in personal development and management. The two main knowledge contributions in this paper are the creation of the first labelled dataset for representational systems, which is now publicly available, and the use of machine learning techniques for the first time to identify a person's preferred representational system in an automated way.

Keywords: machine learning; natural language processing; neuro linguistic programming; representational systems; behavioural patterns

Citation: Amirhosseini, M.H.; Wall, J. A Machine Learning Approach to Identify the Preferred Representational System of a Person. *Multimodal Technol. Interact.* **2022**, *6*, 112. https://doi.org/10.3390/mti6120112

Academic Editor: Mu-Chun Su

Received: 26 October 2022
Accepted: 14 December 2022
Published: 17 December 2022

1. Introduction

Neuro-Linguistic Programming (NLP), created in the early 1970s, is now a popular approach for communication and personal development. It is recognised as a set of techniques which can be used to identify how people think, communicate, and behave [1]. In these techniques, neurological processes, behavioural patterns, and a person's language are used and organized in a certain way to achieve better communication and personal development [2]. NLP provides a collection of techniques, communication tools, and approaches which can be used by people and organisations in order to achieve their goals in personal development and management [3]. NLP has been broadly adopted in different fields, such as education and management, and it has been deployed by well-known companies and organisations such as IBM, NASA, McDonald's, and the U.S. Army [4]. Moreover, NLP is being applied widely, often informally, in the UK higher education system [5] and has become popular with academics and psychologists [6]. Between 2003 and 2010, the UK's education sector employed NLP training in their Fast Track Teacher Programme, in order to train more than 2000 teachers [7]. Additionally, the NHS (National Health Service) in the UK embedded NLP training in more than 300 facilities between

2006 and 2009 to improve the interactions between doctors and patients and facilitate the diagnosis process [7]. Although the application of NLP has had success across different disciplines and confidence in its utility has been increasing [8], few academic works and publications exist on the subject.

NLP asserts that people are innately capable and creative, acting based on how they understand and represent the world, rather than how the world is [2]. In other words, NLP claims that everyone has a mental model according to their experience, beliefs, culture, knowledge, and values. Initially, NLP focused on recognising and understanding the strategy that people use in order to process information. In 1975, Bandler and Grinder, the founders of NLP, published their first NLP book based on the models created by Fritz Prel [9], a well-known psychiatrist, Virginia Stair, a researcher in family therapy, and Milton Erickson, a worldwide recognised psychologist and hypnotherapist. As a result, a roadmap was presented for NLP in order to develop the necessary scientific basis to support its methodology [2]. This then developed into a collection of frameworks, tools, and techniques, which is applicable in different disciplines [10].

1.1. Representational Systems

NLP is used for personal development and includes a variety of techniques. Identification of the preferred representational system of an individual is one of the most important aspects of personal development and NLP techniques can help with this. Representational systems are the different ways that we represent or store information in our mind [11]. This occurs via the five main sensory modalities, seeing, hearing, feeling, tasting, and smelling. People comprehend their environment using these sensory modalities and the information that they receive as a result, which is coded and stored in their mind through these senses, will be filtered with their values and beliefs [12]. As a result, examining representational systems can help us to evaluate and understand how the human mind processes information and interprets meanings [13]. The use of these representational systems is highly dependent on context, varying with the situation [14] and, in specific contexts, a person's mind may prefer to use one or more representational systems to communicate or learn [15,16]. In fact, people use all sensory-based representational systems in different situations, but each person has a dominant preferred representational system, which is used more than other representational systems. This preferred representational system can be used in different ways, such as the way they speak and learn, and in other communicatory pathways [17]. For example, a person may use specific words to describe a situation or may understand something more clearly if some specific words are used in a conversation. Each representational system is associated with specific tendencies of characteristics and the preference of using each of the representational systems in a person can be related to different generalisations of characteristics. Therefore, a lot of information about learning processes, behavioural patterns, and likely characteristics can be revealed through understanding the preferred representational system of an individual [17].

The five representational systems correspond to the main senses, including visual, auditory, kinaesthetic, olfactory, and gustatory (VAKOG). The visual representational system involves the creation of internal images and the observation of things, including pictures, films, charts and diagrams, handouts, and demonstrations [2]. Visual people are interested in what a concept looks like and they usually memorise via observations of imagery [13]. Remembering long verbal instructions can be challenging for them. They also tend to be less distracted by noise. This means that what they see is more important for them and has a priority in comparison to what they experience and understand through hearing, tasting, smelling, or touching.

The auditory representational system involves the comprehension of information through listening [2]. Individuals who prefer this sensory system prioritise their auditory experiences over other senses. Features of this include a greater importance attributed to tone of voice in verbal communication. Moreover, referencing the sounds associated with concepts may be relied on to enhance the conceptualisation of ideas. Individuals learn

and memorise information better through the processing of information via this sensory modality [18]. They are more likely to enjoy music and are, correspondingly, distracted and disturbed by noise more easily.

The kinaesthetic representational system involves internal feelings of physical experience, touching, holding, emotions, and doing practical hands-on activities [2]. People whose preferred representational system is kinaesthetic usually respond very well to touching and physical activities [18]. They are more interested in something that generates a feeling and they usually learn and memorise through doing something [13]. This means that they are less interested in theory and more interested in trying things and doing physical activities.

The olfactory and gustatory representational systems are not considered as main primary representational systems, as they are not popular as a primary representational system for people [19]. As a result, psychologists and NLP practitioners confine their consideration to visual, auditory, and kinaesthetic for assessing the preferred representational system of an individual. It is worth mentioning that each person uses all the sensory modalities in different situations; however, they have a preferred representational system that is used most when speaking, learning, or communicating. As a result, if an individual has one or more missing sensory modalities, they are still using other sensory modalities and their preferred representational system would be one of the remaining sensory modalities that they are using.

1.2. Identification of the Preferred Representation System

NLP experts have recognised that a key identifier of primary dependence on specific sensory modalities is the language that we use. This is because there are different sensory words in language, called 'predicates,' and the use of the related sensory modality can be identified through recognition of these predicates. As an example, when someone says, 'I feel that you are not happy,' the kinaesthetic representational system is involved; however, when someone says, 'I see that you are not happy,' the visual representational system is involved. Accordingly, the language used can indicate the preferred representational system for a person. When a person uses a certain representational system, specific sensory words will be chosen, which can indicate what portion of internal representations the person brings into awareness [20]. This can also facilitate communication, as adopting the language used based on a person's preferred representational system can help them in understanding what one wishes to communicate [21]. The biggest problem in communication is that, when people are listening to you or reading your message, they may not assimilate what is being transmitted [2]. As a result, the identification of predicates and recognising their preferred representational system can be used to improve communication. Table 1 shows examples of key predicates for each representational system.

Table 1. Example of key predicates for each representational system.

Representational System	**Predicates**			
Visual	See	View	Watch	Perspective
	Look	Clear	Image	Light
	Appear	Observe	Vision	Imagine
	Show	Outlook	Picture	Illustrate
	Look	Flash	Sight	Scene
Auditory	Hear	Ring	Talk	Announce
	Listen	Silence	Tell	Outspoken
	Sound	Speechless	Audible	State
	Music	Oral	Voice	Tune in
	Ear	Speak	Echo	Tune out

Table 1. *Cont.*

Representational System	Predicates			
Kinaesthetic	Feel	Push	Flow	Grasp
	Touch	Throw	Heavy	Hard
	Catch	Soft	Rub	Handle
	Hold	Smooth	Solid	Scrape
	Contact	Loose	Shift	Tap

1.3. Background of Automating the Identification Process

There are defined patterns for recognising a person's preferred representational system through analysing conversational language and identifying predicates in sentences used by an individual. However, there are serious challenges involved. Human factors, such as personal judgment, lack of experience, unconscious mistakes, and inaccuracy, can have a direct or indirect impact on the accuracy of the identified preferred representational system. There have been previous studies for improving this accuracy through automating the identification process [8]. However, they can only be considered as a simple computerisation of the process, as no intelligence is involved. Previous methods can be considered as online self-assessment questionnaires providing discrete options to be chosen from, rather than allowing for free expression from the person. As a result, answers are usually based on the available options and the clients' judgment and opinion about themselves. In addition, some of these services need to send the answers to an assessor or NLP practitioner for analysis. In other words, they cannot do a just-in-time analysis and the analysis will be done manually at a later time by a human. Thus, this cannot be called intelligent automation. This can only be considered as the computerisation of the data gathering process. Simplicity is another shortcoming of online surveys, which results in reduced accuracy because of limited considerations. Artificial intelligence and machine learning methods have not been used in the previous attempts at automation in the published literature. Furthermore, there is no available dataset for identification of the preferred representational systems, or any conversational data labelled for each representational system, which could be used for the implementation of such machine learning methods. As a result, this research attempts to develop a unique and comprehensive methodology for creating a labelled dataset and employs machine learning methods for the first time in the prediction of a preferred representational system.

In the following sections, Section 2, 'Methodology,' will cover data collection and labelling strategy, pre-processing and data cleaning, describing training and test sets, and implementation of the machine learning models. Section 3, 'Results and Discussion,' will explain how the implemented models were evaluated and will discuss the confusion matrix, accuracy, F1 score, precision, and recall for each one of the models. Moreover, the performance of 10-Fold cross-validation will be discussed and the performance of the implemented models will be compared. Finally, the conclusions will be provided in Section 4.

2. Methodology

2.1. Data Collection

Until now, there has been no dataset available for identifying a person's preferred representational system, which can be used for machine learning prediction. As a result, a unique approach has been used in this research in order to create such a labelled dataset. After very careful analysis and comparison of the available datasets, the Myers–Briggs personality type dataset [22] was selected. This dataset has been well-used by the research community, including previous work investigating a machine learning approach for personality type prediction [23]. There have not been any associated ethical, legal, or social concerns with this dataset. The original data were collected from the users of an online forum with their consent. The users were asked to complete a questionnaire that recognises their MBTI (Myers–Briggs Type Indicator) type, and then communicate with other users

about personality and different aspects of their life in the online forum via posting texts [24]. The dataset contains 8675 data points, one for each user of the online forum; the attributes of each data point are the MBTI personality type for each person and 50 of the person's posts. This dataset has no specific label for the preferred representational system, rather, the target value was the type of personality.

2.2. Labelling Strategy

A unique methodology was used in this research to identify (label) the preferred representational system for each person in the MBTI dataset, based on the predicates used in their posts, which were obtained from the online forum.

2.2.1. Creating a Collection of Predicates

A collection of relevant vocabulary was created using various relevant documents, including 'Representational Systems' [21], 'The Power of Words' [13], and 'Auspicium NLP Practitioner Home Study Manual (2)' [25]. To do this, five empty lists were created for each representational system, visual, auditory, kinaesthetic, olfactory, and gustatory. The most common predicates were extracted from these documents and each predicate was recorded in the relevant list. The Natural Language Toolkit (LNTK) in Python was used to identify all possible synonyms for each predicate. Five new lists were then created for synonyms of each representational system's predicates and the identified synonyms were recorded in the relevant list. As a result, these 10 lists became a comprehensive collection of all possible predicates for each representational system. All the lists were carefully checked and compared to make sure that there was no overlap. Only a few synonyms were repeated in the different lists, and they were removed to ensure that all the predicates in each list were unique and only related to the corresponding representational system.

2.2.2. Lexical and Syntactic Analysis

After creating a collection of predicates, a lexical and syntactic analysis was done on the Myers–Briggs personality type dataset. The dataset was read row by row and, in each repetition, the 50 posts related to each person were concatenated and then divided into individual sentences to be analysed separately. All these sentences were recorded in a list to be used for lexical and syntactic analysis. Next, the Part-Of-Speech tagging (POS) technique was used in order to identify the role of each word in each sentence. The POS tagger processes a sequence of words and attaches a tag to each word [26]. An example can be seen below, and the meanings of the tags in this example are explained in Table 2.

Input: *'The world is a great place'*
Output:
[('The', 'DT'),
('world', 'NN'),
('is', 'VBZ'),
('a', 'DT'),
('great', 'JJ'),
('place', 'NN')]

There are different POS tag sets which can be used in this process. The Brown corpus [27], one of the most popular POS tag sets, was used in this research. As a result, all nouns, verbs, adverbs, adjectives, and other elements in each sentence were identified.

Table 2. The meanings of tags.

Tag	Meaning
DT	Determiner
NN	Noun, Singular
VBZ	Verb, Present Tense with 3rd Person Singular
JJ	Adjective

A Hidden Markov model (HMM) was used as our tagging technique for building the POS tagger. The aim was to find bigrams, which are two words coming together in the corpus (the entire collection of words/sentences). Thus, a bigram HMM model was created to predict the conditional probability of the next word, based on the assumption that the probability of a word being used depends only on the previous word. The bigram HMM equation is explained below:

$$a_x = argmax_y P(a_y|a_{x-1}, b_y) \quad (1)$$

In order to solve the tagging problem, the nearby words and tags should be checked in this step:

$$a_x = argmax_y P(a_y|a_{x-1}) P(b_y|a_y) \quad (2)$$

In Equation (2), $P(b_y|a_y)$ represents the word likelihood and $P(a_y|a_{x-1})$ represents the tag co-occurrence. Following this step, Equation (3) is used to identify the best sequence of tags:

$$\hat{A} = argmax P(A)P(B|A) \quad (3)$$

Equation (3) can be expanded using the chain rule:

$$P(A)P(B|A) = \prod_{x=1}^{m} P(b_x|b_1a_1 \dots b_{x-1}a_{x-1}a_x)P(a_x|b_1a_1 \dots b_{x-1}a_{x-1}) \quad (4)$$

In the next step, the trigram assumption can be simplified to approximate these two factors. The probability of a word occurring in the sentence only depends on its tag:

$$P(a_x|b_1a_1 \dots a_{x-1}a_1) = P(b_x|a_x) \quad (5)$$

Following this step, the two most recent tags will be used to approximate the probability of the tag:

$$P(a_x|b_1a_1 \dots a_{x-1}) = P(a_x|a_{x-2}a_{x-1}) \quad (6)$$

Finally, the equation can be replaced:

$$P(A)P(B|A) = \prod_{x=3}^{m} P(a_x|a_{x-2}a_{x-1}) \left[\prod_{x=1}^{m} P(b_m|a_m) \right] \quad (7)$$

After applying the POS technique, the stemming technique was used to remove all the prefixes and suffixes, and to identify the root of each word. The reason for doing this is that, in each post, different forms of a word may have been used because of grammatical reasons, and there are also families of derivationally-related words that may have similar meanings [28]. Therefore, it is useful to search for one of these words as a root word, which can be used in future steps for the identification of predicates in the analysed sentences. Thus, all the roots of the words were identified in this step, and they were recorded in a new list, called 'root list,' for the comparison process in the next step.

2.2.3. Comparison Process and Labelling

The POS-tagging list containing the roots of the words was compared with each one of the 10 predicate lists created in Section 2.2.1, containing the predicates for each representational system and the synonyms of the predicates. For each person's 50 posts in the dataset, if any of the words in the 'root list' existed in any of the predicate lists or their synonym lists, a counter for the relevant representational system was incremented; all counters started at 0. After this process, the counter with the highest value represented the preferred representational system and each person in the dataset was labelled with this identified preferred representational system. This process was repeated for every single row in the dataset in order to identify the preferred representational system for all participants and label the dataset with this target value.

In this research, only the three most popular representational systems (visual, auditory, and kinaesthetic) were considered. Any identified olfactory- and gustatory-labelled data were removed from the dataset and not considered for the model training process. As outlined in Section 1.1, the olfactory and gustatory representational systems are not popular and their distribution in the MBTI dataset was not considerable. Figure 1 shows the distribution of Visual, Auditory, and Kinaesthetic samples in the dataset, where 1349 samples (42.81%) were labeled as Visual, 970 samples (30.78%) were labelled as Auditory, and 832 samples (26.40%) were labelled as Kinaesthetic.

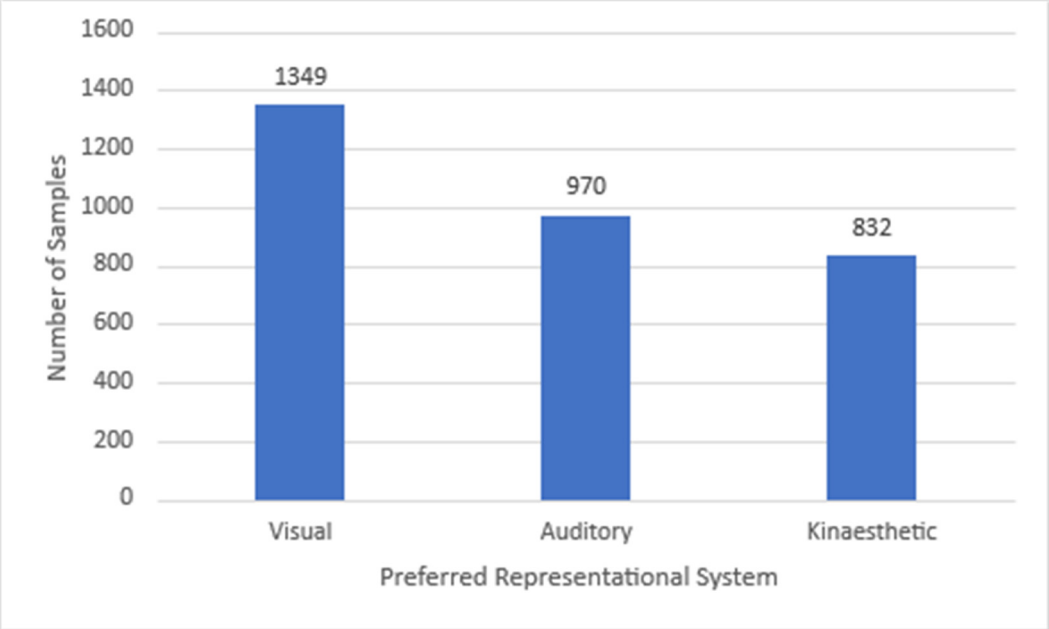

Figure 1. Number of samples for the Visual, Auditory, and Kinaesthetic labels.

2.3. Labelling Validation

In order to evaluate the labelling process, three NLP practitioners were asked to analyse the dataset manually and identify the preferred representational system for each sample based on the number of predicates used in the text. Figure 2 shows the results of their analysis, which shows the number of each representational system identified by the NLP practitioners. The NLP practitioners manually labelled 1354 samples (42.97%) as Visual, 963 samples (30.56%) as Auditory, and 834 samples (26.46%) as Kinaesthetic. The number of preferred representational systems identified manually by NLP practitioners was compared to the number of preferred representational systems identified in Section 2.2.3, and the results are shown in Figure 3.

According to Figure 3, for the visual category, the manual identification was relatively better than the software, where 1354 samples were identified manually and 1349 samples were identified by the software. For the auditory category, 963 samples were identified manually and 970 samples were identified by the software, showing better performance by the latter. For the kinesthetic category, the situation was similar to the visual category in that the manual identification was slightly better than the software, whereby 832 samples were identified manually while 834 samples were identified by the software. Table 3 shows the percentages and the differences between the manual labelling by experts and the automated labelling by the software.

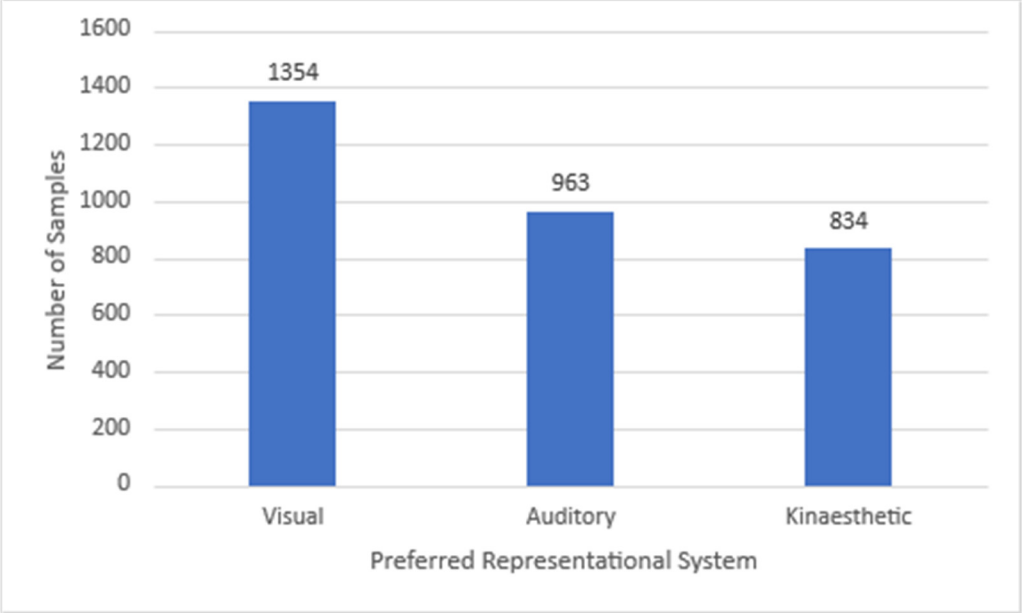

Figure 2. Number of Visual, Auditory and Kinaesthetic preferred representational systems identified by NLP practitioners.

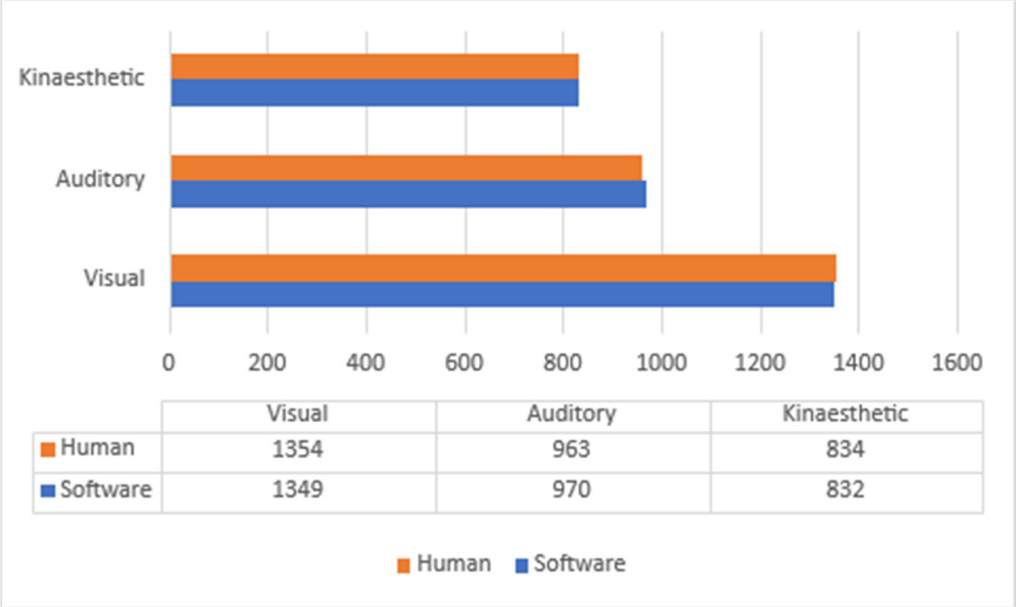

Figure 3. Comparing the number of preferred representational systems identified by human experts and software.

Table 3. Comparing the performance of the human experts and the software.

	Visual	**Auditory**	**Kinaesthetic**
Software	42.81%	30.78%	26.40%
Human	42.97%	30.56%	26.46%
Difference	0.16%	0.22%	0.06

Considering the very small difference between software and human experts, the overall results suggest that the labelling approach in this research is reliable and the algorithm used for labelling is able to replicate the human performance and identify the preferred representational system correctly based on the language used by an individual.

2.4. Pre-Processing and Data Cleaning

During the pre-processing phase, the dataset was analysed for missing data, replacing missing values with 'nan,' In the next step, data cleaning was carried out to make sure that all URLs, punctuations, and stop words were removed, as they do not have any impact on the meaning of the sentence. Following these steps, text vectorization was done in order to convert the text data into a numerical representation and a matrix of TF-IDF (Term Frequency Inverse Document Frequency) features was generated. TF indicates the frequency of each of the words present in the document or dataset, and IDF tells us how important the word is to the context.

The final version of the dataset after data cleaning contains 3151 data points, including the following attributes:

- MBTI Type;
- 50 posts;
- Preferred Representational System of this person.

Table 4 shows which preferred representational system is more common between people with each one of the MBTI personality types.

Table 4. Relationship between MBTI personality types and the preferred representational systems.

MBTI Personality Type	Total Number of Samples	Visual	Auditory	Kinaesthetic
ISTJ	80	**38**	14	28
ISFJ	58	18	**30**	10
INFJ	494	**216**	146	132
INTJ	320	**166**	76	78
ISTP	136	**64**	34	38
ISFP	120	28	**50**	42
INFP	860	**346**	300	214
INTP	398	**182**	120	96
ESTP	38	10	**16**	12
ESFP	14	**8**	0	6
ENFP	254	72	**98**	84
ENTP	214	**110**	42	62
ESTJ	14	**6**	4	4
ESFJ	20	2	**12**	6
ENFJ	62	**34**	14	14
ENTJ	68	**48**	14	6

The train_test_split() function from the sklearn library was used in order to split the dataset into training and testing sets; 70 percent was used for training and 30 percent for testing. The *random_state* parameter was used to initialize the internal random number generator, which decides how the data will be split for training and testing. This was to ensure that the same results would be produced across different runs. This parameter was set to 0 in this research, as we wanted to validate our processing over multiple runs of the code, and we wanted every single data point to be considered during the process every time.

2.5. Machine Learning Models

In this work, as the dataset was labelled and there was a target variable, supervised learning was used for predicting the preferred representational system. Model building and training was carried out for four different supervised machine learning models, Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbor (KNN), using the Scikit-learn library. SVM has recently gained prominence in the field of machine learning and pattern recognition [29] and, using this algorithm, classification is achieved by realising a linear or non-linear separation surface in the input space [30]. For the SVM model, different kernel functions including linear kernel, RBF (Radial Basis

Function) kernel, polynomial kernel, and sigmoid kernel were investigated during the parameter tuning process, and the linear kernel proved to be the optimal kernel function. Textual data, in many dimensions, are usually linearly separable, and the linear kernel will find that separation. Moreover, the *C* parameter (the penalty parameter) instructs the algorithm regarding the false positive classification rate. Different values including 0.1, 1, 10, and 100 for the *C* parameter were evaluated during the parameter tuning process and the best results were achieved when the *C* parameter was set to 100. It was determined that, with a smaller value for this parameter, the classifier misclassifies more data points, because the penalty is so low. Increasing the grid search for this parameter was also considered and a value of 1000 was also investigated; however, it did not provide an optimal result.

Logistic Regression is a method for predicting a dichotomous dependent variable [31], and it uses the maximum-likelihood ratio to determine the statistical significance of the variables [32]. Two hyperparameters, *solver* and *penalty,* were considered for tuning the logistic regression classifier. During the parameter tuning process, different solvers including *newton-cg*, *lbfgs*, *liblinear*, *sag,* and *saga* were investigated for the *solver* hyperparameter and different regularization methods including *none*, *l1*, *l2,* and *elasticnet* were investigated for the *penalty* hyperparameter to investigate which hyper tuning configuration gives the best result. The best results were achieved when the *newton-cg* was used for the *solver* parameter and the *l2* regularization method was used for the *penalty* parameter. Different values including 0.1, 1, 10, and 100 for the *C* parameter were also evaluated during the parameter tuning process, and the best results were achieved when the *C* parameter was set to 100.

The Random Forest algorithm is an ensemble of classification trees, where each tree contributes with a single vote for the assignment of the most frequent class to the input data [33]. During the parameter tuning process for the Random Forest classifier, five parameters including *n_estimators*, *max_features*, *min_sample_leaf*, *random_state,* and *oob_score* were considered in order to optimize the predictive power. The variable *n_estimators* represent the number of trees that the algorithm builds. Increasing the number of trees will lead to a better performance and more stable predictions; however, it can reduce the speed of computation. Different values were tested for this parameter and the best result was achieved when it was set to 1000. The variable *max_features* represents the maximum number of features that the algorithm considers to split a node. There are different options available in the Scikit-learn library for this, and *sqrt* and *log2* were evaluated for this parameter, with *sqrt* giving the best results. The variable *min_sample_leaf* determines the minimum number of end nodes of a decision tree which are required to split an internal node. A smaller number for this parameter makes the model more prone to capturing noise in the training data and provides more reliable results. Different values were tested for this parameter and the best result was achieved when the default value of 1 was used. Finally, the *oob_score* represents the number of correctly predicted data points from the out of bag set. This parameter is used for validating the model and preventing the leakage of data to ensure better performance with low variance. This parameter was set to 'True' for the model developed in this research.

The KNN algorithm has a wide range of applications in the field of machine learning, and it works based on using specific training instances to make a class prediction for a new unclassified instance [34]. Six parameters were considered for tuning the KNN model, *algorithm*, *n_neighbors*, *leaf_size*, *weights,* and *metric*. *Algorithm*, which represents the type of algorithm used to compute the nearest neighbours. Three different algorithms, including *ball_tree*, *kd_tree,* and *brute* were evaluated for this parameter, and the best result was achieved when this parameter was set to *brute*. The *n_neighbors* parameter represents the number of neighbours to be used by default for neighbours queries, and the default value for this parameter is 5. However, different values in a range from 1 to 50 were tested for this parameter and the best result was achieved when this parameter was set to 3. *Leaf_size* can affect the memory required to store the tree and the speed of construction and query. The default value for this parameter is 30. Different values were tested for this parameter

during the tuning process and the best result was achieved when the default value was used. The *weights* parameter represents the weight function used in the prediction. Two possible values for this parameter are *uniform* and *distance*. In the first one, all points in each neighbourhood will be weighted equally and, in the second one, closer neighbours of a query point will have a greater influence than neighbours that are further apart. Both values for this parameter were evaluated and the best result was achieved when this parameter was set to *distance*. Finally, *metric* represents the distance metric used for the tree. The metrics compatible with the brute algorithm are Euclidean, Manhattan, and Minkowski. The best result was achieved when the Minkowski distance metric was used. Table 5 shows a summary of the parameters chosen for each technique. For all four models, all possible combinations of the hyperparameters were investigated during the hyperparameter tuning process and the combinations presented in Table 5 produced the best results.

Table 5. Hyperparameter tuning for each classifier.

Classifier	Parameters	Value
SVM	kernel function	linear kernel
	C	100
Logistic Regression	solver	newton-cg
	penalty	l2 regularization
	C	100
Random Forest	n_estimators	1000
	max_features	sqrt
	min_sample_leaf	1
	random_state	0
	oob_score	TRUE
KNN	algorithm	brute
	neighbors	3
	leaf_size	30
	weights	distance
	metric	Minkowski

3. Results and Discussion

The confusion matrix in Figure 4 visualises the performance of the trained models in this research for the four different model types, SVM, logistic regression, random forest, and KNN. The confusion matrix for these models highlights the multi-class classification of this work, where the target variable has three values: visual, auditory, and kinaesthetic. In the below figure, the columns represent the predicted values of the target variable and the rows represent the actual values of the target variable. In a confusion matrix, true positive (TP) represents the number of predictions when the predicted value matches the actual value, and both are positive. The KNN model has the highest TP value for the visual class. Both logistic regression and SVM have the highest number of correct predictions for the auditory class. SVM has the highest TP value for the kinaesthetic participants.

True negative (TN) in the confusion matrix represents the number of predictions when the predicted value matches the actual value, and both are negative. The SVM model had the highest TN for the visual class; the KNN model had the highest TN for the auditory class; all four models had the same TN for the kinaesthetic class.

False positive (FP) represents a false positive prediction when the actual value was negative, also known as the type 1 error. The highest FP for the visual class was related to the KNN model. The SVM model had the highest FP for the auditory class; the KNN model did not have any FP for this class. The KNN, logistic regression, and SVM models had the same FP for the kinaesthetic class, while the random forest model did not have any FP for this class.

False negative (FN), in the confusion matrix, refers to where the actual value was positive, but the model predicted a negative value, also known as the type 2 error. The

highest FN for the visual class was related to the SVM model; the KNN model did not have any FN for this class. The KNN model has the highest FN for both the auditory and kinaesthetic classes.

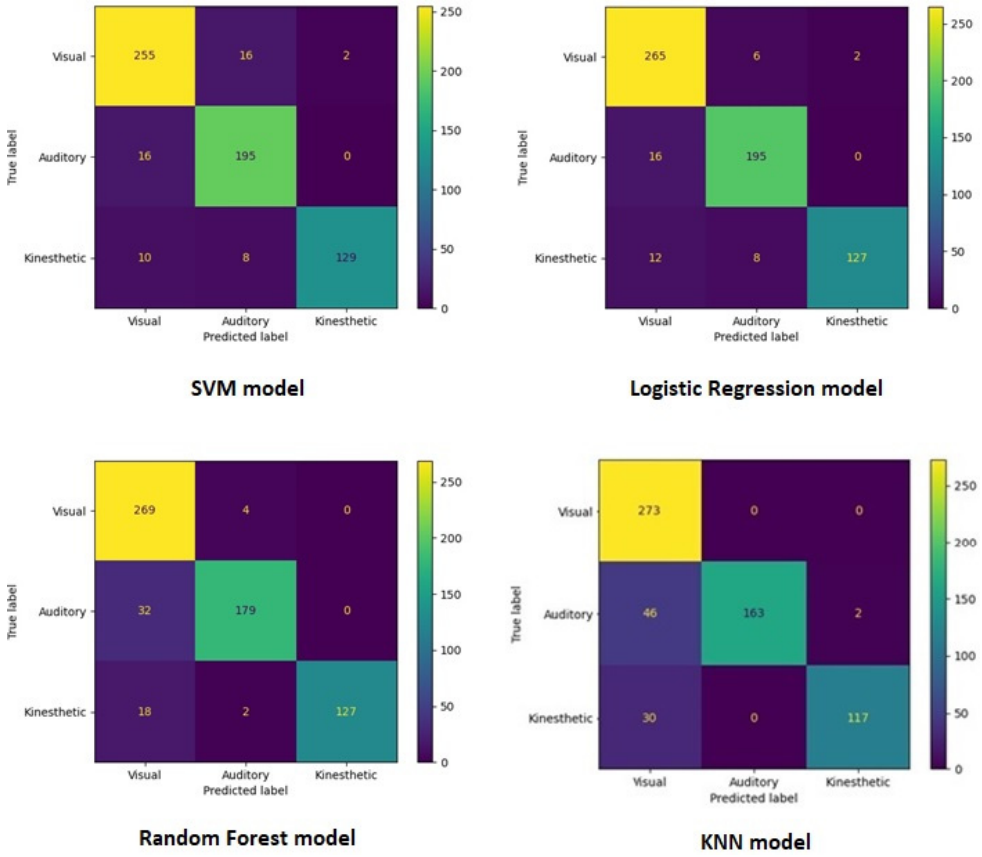

Figure 4. Confusion matrix for SVM, Logistic regression, Random Forest, and KNN models.

Three evaluation metrics including (1) precision, (2) recall, and (3) F1 score can be used in order to evaluate the performance of the models. Precision shows what fraction of correctly predicted samples were positive. The recall evaluation metric shows what fraction of all positive samples were correctly predicted as positive, also known as the probability of detection, sensitivity, or TP rate. Precision and recall can be combined into a single measure, called the F1 score. F1 is the harmonic mean of precision and recall. Table 6 shows the calculated precision, recall, and F1 for each class per model used. Following this step, a micro average precision for each model can be calculated through adding the individual TP, FP, and FN values for the different classes and then applying them to obtain the statistics. In other words, the average of the precision of each model on different classes should be calculated. Table 7 shows the micro average precision for each model. Additionally, the accuracy percentage for each model is presented in Table 8. The logistic regression model had the best performance amongst the four models that were investigated. The SVM and KNN had the same results, while the random forest model showed the weakest performance.

In order to ensure that the results are not biased, the StandardScaler() function from the sklearn library was used to resize the distribution of values in the dataset, with the aim of potential improvement of the performance of the machine learning models. It is possible that the variables in the dataset did not contribute equally to the model fitting and model learning function, and this may have created a bias. Thus, a feature-wise normalisation was used to deal with this potential problem and the models were trained again. In addition, 10-fold cross-validation was performed in order to achieve a more in-depth evaluation of the models. For the cross-validation experiments, accuracy was reported as the evaluation

metric, so that the accuracy percentage could be compared with the presented results in Table 8. After calculating the accuracy score for each fold, the mean classification accuracy on the dataset was calculated. Table 9 shows the accuracy scores calculated for each fold and the mean classification accuracy as the final accuracy score for each model.

Table 6. Precision, Recall, and F1 score for each class and model.

Classifier	Class	Precision	Recall	F1 Score
SVM	Visual	0.91	0.93	0.92
	Auditory	0.89	0.92	0.91
	Kinaesthetic	0.98	0.88	0.93
Logistic Regression	Visual	0.9	0.97	0.94
	Auditory	0.93	0.92	0.93
	Kinaesthetic	0.98	0.86	0.92
Random Forest	Visual	0.84	0.99	0.91
	Auditory	0.97	0.85	0.9
	Kinaesthetic	1	0.86	0.93
KNN	Visual	0.78	1	0.88
	Auditory	1	0.77	0.87
	Kinaesthetic	0.98	0.8	0.88

Table 7. Micro average Precision, micro average Recall, and micro average F1 score for each model.

Classifier	Micro Average Precision	Micro Average Recall	Micro Average F1 Score
SVM	0.93	0.91	0.92
Logistic Regression	0.94	0.92	0.93
Random Forest	0.94	0.90	0.91
KNN	0.92	0.86	0.88

Table 8. Accuracy for each model.

Classifier	Accuracy Percentage
SVM	91%
Logistic Regression	93%
Random Forest	91%
KNN	87%

According to Table 9, the overall performance of the SVM, Random Forest, and KNN models improved, and the performance of the Logistic Regression model remained the same. Table 9 also shows the comparison between the 10-Fold cross-validation results and the accuracy scores from Table 8.

Considering the mean accuracy score as the final accuracy score after performing 10-fold cross-validation, Table 10 shows that the SVM model had the best performance with an accuracy of 96%. Random Forest was the second model with an accuracy of 95%. KNN and Logistic Regression, with accuracies of 95% and 93%, respectively, were the third- and fourth-best performing models. It should also be mentioned that using the StandardScaler() function and normalising the data had a significant impact on the performance of the models, resulting in a 5% improvement in accuracy score for the SVM model, 4% improvement for the Random Forest model, and 7% improvement for the KNN model. Furthermore, we can confidently claim that the results are not biased as the dataset has been normalised and the performance of the models has been evaluated in 10

different iterations using different parts of the dataset with the same size at each step of the validation process.

Table 9. 10-Fold cross-validation results for each classifier.

Classifier	Fold Number	Accuracy Scores Calculated for Each Fold	Mean Accuracy Score
SVM	1	0.96835443	0.96
	2	0.95555556	
	3	0.96190476	
	4	0.96825397	
	5	0.94920635	
	6	0.96825397	
	7	0.94285714	
	8	0.96190476	
	9	0.94920635	
	10	0.95555556	
Logistic Regression	1	0.95253165	0.93
	2	0.93650794	
	3	0.93333333	
	4	0.91428571	
	5	0.92380952	
	6	0.93015873	
	7	0.90793651	
	8	0.95238095	
	9	0.91746032	
	10	0.92698413	
Random Forest	1	0.9556962	0.95
	2	0.94920635	
	3	0.95555556	
	4	0.94920635	
	5	0.94920635	
	6	0.96190476	
	7	0.94920635	
	8	0.94920635	
	9	0.94285714	
	10	0.94920635	
KNN	1	0.94936709	0.94
	2	0.93650794	
	3	0.95555556	
	4	0.94920635	
	5	0.93650794	
	6	0.94920635	
	7	0.94285714	
	8	0.94920635	
	9	0.93650794	
	10	0.94920635	

Table 10. Accuracy for each model.

Classifier	Accuracy Percentage	Mean Accuracy Score after 10-Fold Cross-Validation
SVM	91%	96%
Logistic Regression	93%	93%
Random Forest	91%	95%
KNN	87%	94%

4. Conclusions

This research has proposed a novel methodology to create a labelled dataset for the preferred representational system of people. There has been no such dataset available before and this dataset can be used for training machine learning models for the prediction of their preferred representational system. The algorithm used for data labelling was evaluated through a robust comparison between the labels produced by human experts and the software. The difference for all labels, Visual, Auditory, and Kineasthetic, was less than 0.3%, which shows that the labelling approach in this research was reliable. Based on this dataset, a machine learning approach has been investigated to identify the preferred representational system of a person. Four machine learning models including SVM, Logistic Regression, Random Forest, and KNN were trained and compared. A confusion matrix and a range of evaluation metrics were used to analyse the performance of these models. The results show that the Logistic Regression model had the best performance, with 93% accuracy when the dataset was split in the ratio of 70:30. However, after this step, feature-wise normalisation and 10-fold cross-validation were performed to improve the performance of the models and to make sure that the results were not biased. A mean accuracy score was calculated for each model and considered as the final accuracy score. The results show that feature-wise normalisation had a significant impact on the performance of the models, and the SVM, with 96%, accuracy had the highest accuracy score of all four models.

Regarding the knowledge contribution of this paper, this is the first time that machine learning has been used to predict the preferred representational system of people. Moreover, the first labelled dataset for representational systems has been created, and it is now publicly available. The presented methodology in this research can effectively assist NLP practitioners and psychologists to identify the preferred representational system of a person and the relevant cognitive processes. The output of this research can also be helpful for managers in organisations, as it can facilitate the process of improving communication and performance. Moreover, it can be useful for businesses to improve their sales and marketing, as communication has a very strong impact on these.

Author Contributions: Conceptualization, M.H.A.; methodology, M.H.A.; implementation, M.H.A.; validation, M.H.A. and J.W.; formal analysis, M.H.A. and J.W.; data curation, M.H.A.; visualization, M.H.A.; writing—original draft preparation, M.H.A.; writing—review and editing, M.H.A. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available at: https://drive.google.com/file/d/1FFl2Mwfe04WKB0ccgog7Z5y9ku80yRV2/view?usp=sharing (accessed on 15 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Transform Destiny (2015) Introduction to Neuro-Linguistic Programming. Available online: https://www.freenlphomestudy.com/membersonly/iNLP/iNLPManual.pdf (accessed on 12 May 2022).
2. Júnior, M.C.; Mendonça, M.; Farias, M.A.D.F.; Henrique, P.; Corumba, D. A Neurolinguistic Method for Identifying OSS Developers' Context-Specific Preferred Representational Systems. In Proceedings of the Seventh International Conference on Software Engineering Advances, Lisbon, Portugal, 18–23 November 2012; Available online: https://www.researchgate.net/publication/327561064_A_Neurolinguistic_Method_for_Identifying_OSS_Developers\T1\textquoteright_Context-Specific_Preferred_Representational_Systems (accessed on 12 May 2022).
3. Lazarus, J. *Successful NLP: For the Results You Want*; Crimson Publishing: Surrey, UK, 2010.
4. Witkowski, T. Thirty-five years of research on neuro-linguistic programming. NLP research data base. State of the art or pseudoscientific decoration? *Pol. Psychol. Bull.* **2010**, *41*, 58–66. [CrossRef]

5. Singer, M.T.; Lalich, J. *'Crazy Therapies'. What Are They? Do They Work?* Jossey-Bass: San Francisco, CA, USA, 1996.
6. Tosey, P.; Mathison, J. Neuro-linguistic programming and learning theory: A response. *Curric. J.* **2003**, *14*, 371–388. [CrossRef]
7. Kotera, Y. NLP: Why This Therapy Has Failed to Join the Mainstream. 2019. Available online: https://www.derby.ac.uk/blog/neuro-linguistic-programming-psychology (accessed on 10 October 2022).
8. Amirhosseini, M.H.; Kazemian, H. Automating the process of identifying the preferred representational system in Neuro Linguistic Programming using Natural Language Processing. *Cogn. Process.* **2019**, *20*, 175–193. Available online: https://link.springer.com/article/10.1007/s10339-019-00912-3 (accessed on 12 May 2022). [CrossRef] [PubMed]
9. Clayton, M. John Grinder and Richard Bandler: NLP (Neuro Linguistic Programming). 2017. Available online: https://www.pocketbook.co.uk/blog/2017/08/01/john-grinder-and-richard-bandler-nlp-neuro-linguistic-programming (accessed on 10 October 2022).
10. Tosey, P.; Mathison, J. *Introducing Neuro-Linguistic Programming Centre for Management Learning & Development*; School of Management, University of Surrey: Guildford, Australia, 2006.
11. Ellerton, R. NLP's Auditory Digital Representational Systems. 2007. Available online: http://www.renewal.ca/nlp48.html (accessed on 12 May 2022).
12. O'Connor, J.; Seymour, J. *Introducing Neuro Linguistic Programming: Psychological Skills for Understanding and Influencing People*, Revised ed.; The Aquarian Press: London, UK, 1993.
13. McAfee, K. 'The Power of Words', an Introduction to NLP Representational Systems, Becoming a More Effective Communicator. 2009. Available online: https://pdfcoffee.com/nlp-representational-systemspower-of-wordspdf-pdf-free.html (accessed on 12 March 2022).
14. Einspruch Eric, L.; Forman Bruce, D. Observations concerning Research Literature on Neuro-Linguistic Programming. *J. Couns. Psychol.* **1985**, *32*, 589–596. [CrossRef]
15. Matthews, D.B. Learning Styles Research: Implications for Increasing Students in Teacher Education Programs. *J. Instr. Psychol.* **1991**, *18*, 228–236.
16. Peters, D.; Jones, G.; Peters, J. Preferred 'learning styles' in students studying sports-related programme in higher education in the United Kingdom. *Stud. High. Educ.* **2008**, *33*, 155–166. [CrossRef]
17. NLP Dynamics Ltd. Representational Systems. 2013. Available online: http://www.distancelearning.academy/wp-content/uploads/2015/02/Representational-Systems.pdf (accessed on 13 March 2022).
18. Bensted, C. Representational Systems. 2014. Available online: http://badis.co.uk/resources/Repsys.pdf (accessed on 12 May 2022).
19. Rayner Institute. The Representational Systems. 2015. Available online: http://www.raynerinstitute.com/uploads/9/8/6/1/98611170/nlp_rep_system.pdf (accessed on 12 May 2022).
20. Dilts, R.; Grinder, J.; Bandler, R.; DeLozier, J. *Neuro-linguistic Programming: Volume 1*; Meta Publications: Cupertino, CA, USA, 1980.
21. Brefi Group Limited Representational System. 2004. Available online: https://www.scribd.com/document/230814203/Nlp-Representational-Systems-Test (accessed on 14 March 2022).
22. Kaggle. Myers-Briggs Personality Type Dataset. Includes a Large Number of People's MBTI Type and Content Written by Them. 2021. Available online: https://www.kaggle.com/datasnaek/mbti-type (accessed on 12 May 2022).
23. Amirhosseini, M.H.; Kazemian, H. Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®. *Multimodal Technol. Interact.* **2020**, *4*, 9. [CrossRef]
24. Hernandez, R.; Knight, I.S. Predicting Myers-Bridge Type Indicator with text classification. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017. Available online: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6839354.pdf (accessed on 12 May 2022).
25. Auspicium Limited Auspicium NLP Practitioner Home Study Manual (2). 2012. Available online: https://toaz.info/doc-viewer (accessed on 23 March 2022).
26. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media Inc.: Newton, MA, USA, 2009.
27. W3-Corpora Project The Brown Corpus. 1998. Available online: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html (accessed on 15 May 2022).
28. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2009.
29. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2000.
30. Vishwanathan, S.V.M.; Murty, M.N. SSVM: A simple SVM algorithm. In Proceedings of the 2002 International Joint Conference on Neural Networks, Honolulu, HI, USA, 12–17 May 2002; Volume 3, pp. 2393–2398. [CrossRef]
31. Kurt, I.; Ture, M.; Kurum, A.T. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst. Appl.* **2008**, *34*, 366–374. [CrossRef]
32. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley & Sons: New York, NY, USA, 2000.
33. Rodriguez-Galiano, V.F.; Chica-Olmo, M.; Abarca-Hernandez, F.; Atkinson, P.M.; Jeganathan, C. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sens. Environ.* **2012**, *121*, 93–107. [CrossRef]
34. Larose, D.T. *Discovering Knowledge in Data*; Wiley & Sons Inc.: Hoboken, NJ, USA, 2005; pp. 96–103.