BioMedInformatics

MDPI

Article

Ensemble Methods to Optimize Automated Text Classification in Avatar Therapy

Alexandre Hudon [1,2], Kingsada Phraxayavong [3], Stéphane Potvin [1,2] and Alexandre Dumais [1,2,3,4,*]

[1] Department of Psychiatry and Addictology, Faculty of Medicine, Université de Montréal, Montreal, QC H3T 1J4, Canada; alexandre.hudon.1@umontreal.ca (A.H.); stephane.potvin@umontreal.ca (S.P.)
[2] Centre de Recherche de l'Institut Universitaire en Santé Mentale de Montreal, Montreal, QC H1N 3J4, Canada
[3] Services et Recherches Psychiatriques AD, Montreal, QC H1C 1H1, Canada; kingsada@me.com
[4] Institut National de Psychiatrie Légale Philippe-Pinel, Montreal, QC H1C 1H1, Canada
* Correspondence: alexandre.dumais@umontreal.ca

Abstract: Background: Psychotherapeutic approaches such as Avatar Therapy (AT) are novel therapeutic attempts to help patients diagnosed with treatment-resistant schizophrenia. Qualitative analyses of immersive sessions of AT have been undertaken to enhance and refine the existing interventions taking place in this therapy. To account for the time-consuming and costly nature and potential misclassification biases, prior implementation of a Linear Support Vector Classifier provided helpful insight. Single model implementation for text classification is often limited, especially for datasets containing imbalanced data. The main objective of this study is to evaluate the change in accuracy of automated text classification machine learning algorithms when using an ensemble approach for immersive session verbatims of AT. Methods: An ensemble model, comprising five machine learning algorithms, was implemented to conduct text classification for avatar and patient interactions. The models included in this study are: Multinomial Naïve Bayes, Linear Support Vector Classifier, Multi-layer perceptron classifier, XGBClassifier and the K-Nearest-Neighbor model. Accuracy, precision, recall and f1-score were compared for the individual classifiers and the ensemble model. Results: The ensemble model performed better than its individual counterparts for accuracy. Conclusion: Using an ensemble methodological approach, this methodology might be employed in future research to provide insight into the interactions being categorized and the therapeutical outcome of patients based on their experience with AT with optimal precision.

Keywords: virtual reality therapy; artificial intelligence; auditory hallucinations; schizophrenia; psychotherapy; machine learning; ensemble modeling; text classification

Citation: Hudon, A.; Phraxayavong, K.; Potvin, S.; Dumais, A. Ensemble Methods to Optimize Automated Text Classification in Avatar Therapy. *BioMedInformatics* **2024**, *4*, 423–436. https://doi.org/10.3390/biomedinformatics4010024

Academic Editors: Alexandre G. De Brevern and Jörn Lötsch

Received: 30 August 2023
Revised: 15 January 2024
Accepted: 5 February 2024
Published: 7 February 2024

1. Introduction

Schizophrenia is a complex psychopathology characterized by positive symptoms (such as auditory hallucinations, persecutory delusions, disorganization of thoughts and behaviors) and negative symptoms (such as avolition, anhedonia, alogia) [1,2]. Pharmacological treatment of schizophrenia focuses primarily on positive symptoms because they can be linked to serious deleterious events (such as suicide and violence) [3,4]. However, recent studies have demonstrated that around 25 to 30% of patients are resistant to current lines of treatment [5–7]. Multiple definition exists when referring to treatment resistance [8]. The most common accepted definition is when after two trials with antipsychotic medicines with verified adherence and an appropriate dose and duration, symptoms persist [9]. Patients who meet this definition are known as patients with treatment-resistant schizophrenia (TRS). Clozapine, a second-generation antipsychotic, is commonly used to treat patients with TRS [10]. Up to 60 percent of patients taking clozapine respond poorly to this approach, which is why further approaches, notably adjunct to medication, have been used or are currently being studied [11].

Once such approach is psychotherapy. Amongst psychotherapeutic approaches, cognitive behavioral therapy (CBT) is one of the most used [12]. Despite statistical improvements of patients, little evidence has been found to recommend its use routinely in patients with TRS considering the lack of evidence in clinical improvements [13]. Therefore, further therapeutical approaches are currently being studied, such as Avatar Therapy (AT) [14]. The effectiveness of this treatment in lowering patients' resistant auditory hallucinations and gauging their wellbeing is still being investigated [15]. In order to interact with patients in an immersive setting, AT recommends utilizing a virtual reality headset [16]. In AT, a therapist simulates a patient's auditory hallucinations while using the virtual environment by the means of an animated visual depiction (pre-configured in collaboration with the patient). The Leff et al. (2014) team in London, United Kingdom, created AT in 2008 [17]. The first randomized controlled, single-blinded, trial, was conducted in South London and Maudsley NHS Trust (United Kingdom) in 2016 with 150 adult patients who had been clinically diagnosed with schizophrenia spectrum and continued to experience auditory verbal hallucinations despite receiving treatment [18]. AT or supportive therapy was randomly assigned to these patients. Evaluated by the Psychotic Symptoms Rating Scales Auditory Hallucinations (PSYRATS-AH), the primary result was a decrease in auditory verbal hallucinations, as well as a decrease in depressive symptoms at 12 weeks [18]. A current clinical trial at the University Institute in Mental Health of Montreal (IUSMM) that compares CBT to AT for patients with schizophrenia who are receiving ongoing therapy and experiencing auditory hallucinations is currently taking place. In this study comprising 136 patients, 68 are receiving AT, and 68 are receiving CBT. While this experiment is being conducted, 37 patients who participated in AT and 37 who participated in CBT were evaluated during a 365-day pilot randomized comparison trial at the IUSMM to determine the efficacy of AT over CBT for this population. For these individuals, AT performed better than CBT for auditory hallucinations, and it also significantly improved the quality of life and persecutory beliefs [19].

To assess the content of AT and provide a more comprehensive grasp of the dynamics taking place between a patient and their avatar during immersive sessions, qualitative analyses have been conducted. A preliminary content analysis of AT was performed in 2018 by Dellazizzo et al., who explored the treatment of 12 AT patients. A total of five themes emerged from patients' conversations with the avatar, according to this analysis: emotional response to voices, ideas about voices and schizophrenia, oneself, coping techniques, and goals [20]. This analysis provided the first insight into prospective AT treatment targets. In a follow-up study in 2021, Beaudoin et al. qualitatively evaluated 125 therapy verbatims of patients who received AT. The avatar's two main key interaction themes were confrontational techniques (which had eight sub-themes) and positive techniques (which included six sub-themes). The patients' emotional reactions, self-perceptions, coping strategies, goals and notions of voices and schizophrenia were all highlighted as five key themes. A total of 14 sub-themes was identified amongst these 5 main themes [21]. By illuminating the interactions between avatars and patients, it was possible to highlight important areas of focus that may direct future research and therapeutic interventions, these descriptive investigations advance our understanding of the therapeutic process of AT. While descriptive data may offer extensive insights, they lack the quantitative counterpart necessary for identifying the precise components of psychotherapy that may help patients achieve favorable outcomes [22]. Qualitative analyses are also costly and time-consuming and are subject to inherent biases such as misclassification bias, which is even more prevalent when different kinds of interaction can overlap (which is frequent in natural language) [23–25].

To provide a quantitative propensity for qualitative analyses of such verbatims, classification techniques can be used [26]. This is usually carried out using machine learning algorithms. Classification techniques can be supervised (data are deduced from a labeled dataset) or unsupervised (data are inferred) [27]. One major problem is that often, such implementations need large datasets to derive accurate classification [28]. Another lim-

itation is the limited data availability in the psychotherapeutic setting considering the confidential nature of interactions between patients and their therapists. A recent literature review on machine learning algorithms with small datasets in the context of psychotherapy identified several key algorithms that can perform acceptably on such datasets [29]. The first implementation of such a technique on AT verbatims was carried out using a linear vector classifier (LSVC), as per this literature review, and concluded that it can conduct automated theme classifications in AT session transcripts using a limited dataset, achieving accuracy and substantial classification agreement comparable to that of human coders [30]. This technique was also found to be useful in efficiently identifying interactions between the avatar and the patient in AT [31]. However, this approach is limited by the linear assumptions of LSVC (i.e., interactions in AT are assumed to be entirely linearly separable), by its sensitivity to data outliers and by the difficulty in successfully classifying imbalanced data [32]. The AT dataset contains imbalanced data due to the fact that some types of interactions between patient and avatar are more frequent.

To account for single model limitations, ensemble modeling is a technique that is widely used [33]. It consists of creating a better, more precise, and more reliable predictive model by combining the predictions of various distinct models [34]. This increases the overall effectiveness and generalization of the ensemble by making use of the diversity of predictions made by the individual models [35]. Such an approach can increase the model's complexity and the computational resources needed for its performance [33]. However, considering the small datasets employed for automatic classification in AT, these limitations are insignificant compared to the expected potential. To our knowledge, this has never been conducted for psychotherapeutic content and was never attempted on AT verbatims.

The main objective of this study is to assess the change in accuracy of automated text classification machine learning algorithms when using an ensemble approach on immersive session verbatims of AT. It is hypothesized that such an approach will increase the accuracy of automated text classification in AT and will therefore yield better automated text classification for future analyses of verbatims for patients who are receiving AT. When taking into account the large number of variables being incorporated into the automated classification of the interactions occurring in the verbatims for AT, a combination of different machine learning classification models is believed to account for potential misclassification as compared to the use of a single classification model.

2. Materials and Methods

2.1. Recruitment and Participants

The dataset utilized in this investigation consists of therapeutic interactions of participants involved in a pilot trial carried out at the Research Center of the University Institute in Mental Health of Montreal (CR-IUSMM). They were all affected by treatment-resistant schizophrenia (TRS), which is marked by continuous auditory hallucinations despite the use of two or more dopaminergic antagonists. Their AT sessions were conducted between 2017 and 2022. The clinical trial can be found on ClinicalTrials.gov under the identifier NCT03585127 [19]. Each participant underwent a series of nine one-hour psychotherapeutic sessions, with eight of them being immersive sessions that included interaction with a virtual representation of their auditory verbal hallucinations known as the avatar. The study comprised individuals who were over 18 years old and were patients at the IUSMM.

2.2. Dataset: Corpus of Avatar Therapy and Features

Research assistants transcribed verbatim the immersive sessions of 18 AT patients from audio recordings. Subsequently, AH reviewed the verbatims to ensure transcription accuracy, yielding a total of 144 verbatims, representing nearly 70 h of AT immersion. Interactions between patients and avatars were annotated and categorized based on the 27 themes specified in Beaudoin et al. (2021) [21]. Table 1 presents the categorized interaction themes for both the avatar and the patients.

Table 1. Themes and samples of interactions between avatars and patients as outlined by Beaudoin et al. (2021) [21].

Avatar Themes	Samples	Patient Themes	Samples
Accusations	"You've carried out this task."	Approbation	"Your observation is accurate." "I'm capable of achieving this."
Omnipotence	"I'm feeling scattered everywhere."	Self-deprecation	
Beliefs	"In my opinion, your behavior seems irrational."	Self-appraisal	"I consider myself a kind individual."
Active listening, empathy	"Take your time to unwind, please."	Other beliefs	"You're the one with control over me."
Incitements, orders	"I recommend discontinuing this activity."	Counterattack	"You're responsible for this, not me!"
Coping mechanisms	"Can you explain why my mentioning this makes you sad?"	Maliciousness of the voice	"You seem to be intentionally complicating things for everyone."
Threats	"I'll bring about your downfall."	Negative	"This is quite challenging."
Negative emotions	"Coming to terms with that is challenging for me."	Negation	"I didn't perform this action."
Self-perceptions	"I view myself as being insignificant."	Omnipotence	"I possess unmatched abilities."
Positive emotions	"I'm unparalleled in the entire world."	Disappearance of the voice	"Please disappear!"
Provocation	"Try preventing me from causing you harm."	Positive	"I'm experiencing a positive emotional state."
Reconciliation	"Shall we work towards reconciliation?"	Prevention	"I'll attempt to ignore your presence."
Reinforcement	"Give this another attempt."	Reconciliation of the voice	"Shall we become friends?"
		Self-affirmation	"I am capable of accomplishing this."

A dataset was created using 144 therapy transcripts from 18 randomly chosen patients who received AT at the CR-IUSMM between 2017 and 2022. Eight treatment sessions were attended by each patient, resulting in an average of eight transcripts per subject. The initial transcripts were meticulously typed in Canadian French. To annotate the transcripts, the twenty-seven themes listed in the study by Beaudoin et al. (2021) were applied manually [21]. The software QDA Miner version 5, developed by Provalis Research, was utilized for qualitative data analysis for the annotation process [36]. The annotations were then retrieved into text files, each of which contained one to forty interactions that were all related to the same theme. According to the model depicted in Figure 1, the annotations retrieved were subsequently separated into two conceptual databases: one for the avatar and one for the patient. The different themes found in the dataset were balanced.

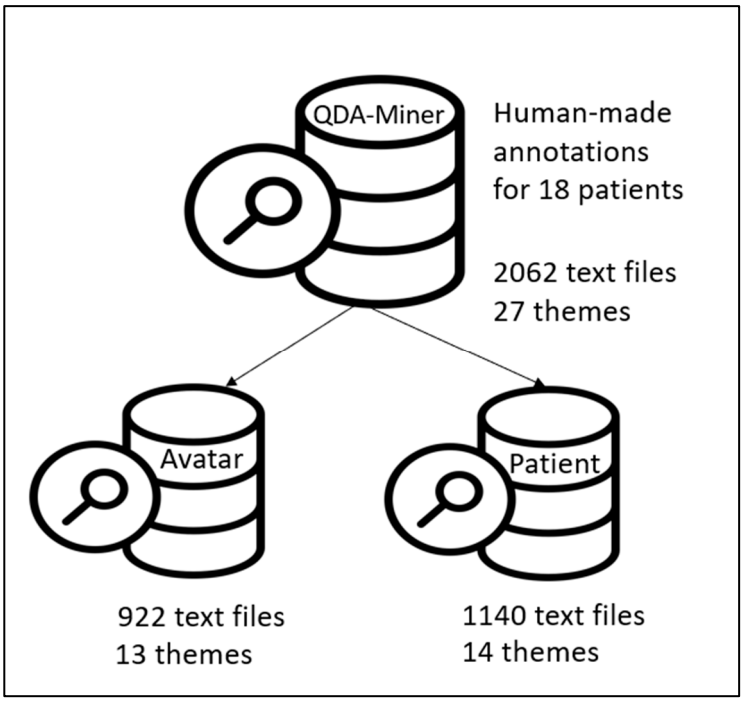

Figure 1. Dataset for the Avatar Therapy corpus.

2.3. Ensemble Modeling

Ensemble modeling implies the use of several classification models to select the best-performing model according to a vote for each classification conducted. In this study, the models being implemented as part of the ensemble are as follows: LSVC, Multinomial Naïve Bayes (Multinomial NB), Multi-layer perceptron classifier (MLP), XGBClassifier (XGB) and the K-Nearest-Neighbor (KNN) algorithms. These were selected based on the previous literature review on the best-performing algorithms for small datasets and the ground rules for composing an ensemble model, notably using diverse base models, the potential for cross-validation and avoidance of highly correlated models. The ensemble model is presented in Figure 2. Ensemble modeling functions were selected and used from the Scikit-Learn library with Python 3.11 [37].

Both the patient conceptual dataset and the avatar conceptual dataset were employed by the ensemble model. To refine classification techniques and optimize the machine learning algorithm's performance, a GridSearchCV (GSCV) approach from the Scikit-Learn library was implemented [37]. Users can explore various hyperparameters and cross-validate the classifier's predictions using the GSCV tool to identify the optimal set of parameters that yield the highest performance. In this study, LSVC classifiers were both subjected to GSCV. The MLP, Multinomial NB classifiers and XGB performed better when considering hyperparameterization; hence, default values were used for these. The KNN was initialized with a default Minkowski distance of three, which is consequent with previous instantiation and analysis of KNN performances on AT [38].

The term frequency–inverse document frequency (TF-IDF) method, which outperforms other algorithm–tokenizer combinations in text categorization, was used in conjunction with the algorithms [39]. The implementation of TfidfVectorizer, offered by the Scikit-Learn module, to implement TF-IDF tokenization, was used [37]. The raw text retrieved from the therapeutic interactions between the avatars and patients during immersive sessions were converted into numerical vectors.

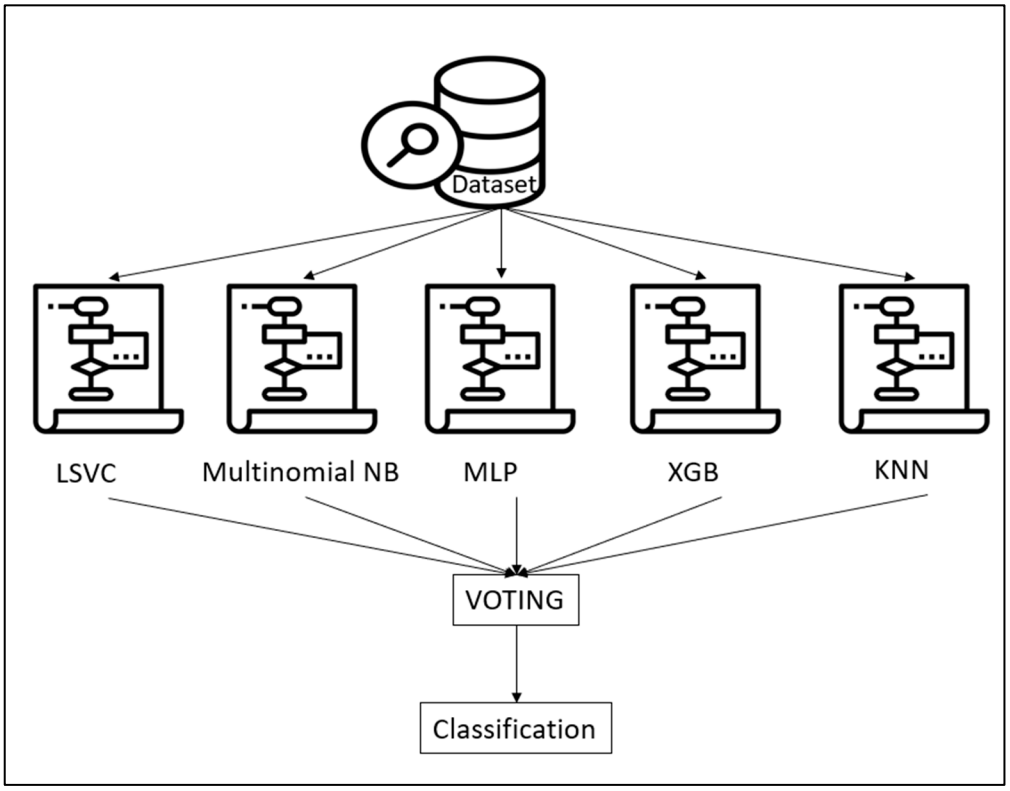

Figure 2. Ensemble modeling is applied to each classification, and this process carried out for both the avatar conceptual dataset and the patient conceptual dataset.

2.4. Machine Learning Algorithms

The five models used to compose the ensemble model are listed below.

1. LSVC: The Support Vector Machine (SVM) approach aims to determine the optimal hyperplane for dividing various classes of data points in a high-dimensional feature space. This involves maximizing the margin between classes to achieve robust generalization performance. The method identifies support vectors, a subset of training samples serving as pivotal points for the decision boundary. Unlike SVC, the LSVC uses a linear kernel. A kernel is a mathematical function transforming data into a higher-dimensional feature space, crucial in handling complex problems that may be challenging or impossible in the original input space. A linear kernel is applied when data separation can be achieved linearly. The implementation of SVC in this study utilized Scikit-Learn, specifically the SVC class from the SVM library [37].
2. The Multinomial Naive Bayes method is a derivative of the Naïve Bayes technique, which, based on the Bayes theorem, assumes conditional independence of features given the class. This method is developed using the Bayes theorem, which enables the updating of the probability of Event A occurring in light of new information or additional supporting evidence from Event B. It calculates the posterior probability ($P(A \mid B)$) by combining the prior probability ($P(A)$) and the likelihood ($P(B \mid A)$). Specifically designed to handle discrete features in text data, such as word counts or frequencies, the Multinomial Naive Bayes classifier is implemented using Scikit-Learn, with the Multinomial NB class from the naive_bayes module employed in this study [37].
3. MLP: The Multi-Layer Perceptron (MLP) classifier serves for classification and various machine learning applications. It constitutes a feedforward neural network model characterized by multiple layers of interconnected neurons. The typical structure of the MLP classifier includes an input layer, one or more hidden layers, and an output layer. Each layer comprises numerous neurons that conduct computations

on incoming data and transmit the results to the subsequent layer. In an MLP, every neuron in each layer is connected to every other neuron in both the layers above and below, indicating full connectivity. The strengths and relevance of information flow across the network are influenced by weights associated with neuron connections. In this study, the MLP implementation is derived from Scikit-Learn, specifically utilizing the MLPClassifier class from the neural_network library [37].

4. XGB: XGB works by sequentially iteratively generating an ensemble of weak learners such as decision trees. Each successive model is then used to correct the mistakes of the prior ones. This is performed to optimize the ensemble's predictive capability while reducing overfitting by minimizing a user-defined loss function. XGboost library, namely the XGBoost class of the XGBClassifier module, provided the XGB implementation used in this study [37].
5. KNN: This widely employed technique is based on the principles of k-means clustering, aiming to group similar data in a comprehensible, relatively swift, and scalable manner while ensuring convergence. It assesses whether two items are identical and organizes them based on their Euclidean distance, representing the length of a line drawn between two data points. The number of clusters is predetermined, and the process unfolds iteratively. Beginning with the random selection of the center (centroid) for each cluster, the Euclidean distance from all data points to the centroids is computed, and the data points are assigned to their closest clusters. Subsequently, for each cluster, a new centroid is determined by calculating the mean of all data points within the cluster. This process is repeated until all points converge, and the cluster centers cease to move. The KNN implementation in this study is sourced from Scikit-Learn, specifically utilizing the neighbors class from the SVM library [37].

2.5. Voting Technique

Ensemble modeling offers different avenues to compare predictions of the classification models they encompass. This technique is useful as classification depends on the performance of multiple models and will therefore not be hampered by big errors or misclassifications from a single model. A bad performance from one model can be compensated for by a great performance from another. One of the most widely used techniques to assess the accuracy of ensemble models is the voting process [40].

There exist two categories of voting techniques: hard and soft voting [41]. In voting techniques applied to AT, hard voting implies adding up all the forecasts for each interaction theme and predicting the interaction theme with the most votes [42]. Soft voting involves summing the anticipated probabilities (accuracy scores) for each interaction theme and predicting the theme with the highest probability [43]. Hard voting is useful when the models in the ensemble are diverse and do not give well-calibrated probabilities, as compared to soft voting (which is better at capturing the nuances of different models' confidence levels) [41]. Considering that nuances of the different models' confidence levels are needed as data cannot be assumed to be well balanced across the different interaction themes, soft voting is used in this study.

2.6. Data Analysis and Validation

Data regarding the classification performance of each theme, encompassing the recall, accuracy, and f1-Score for each algorithm, were compiled using the Classification Report tool within the Scikit-Learn metrics module [37]. The overall average accuracy is established by the ratio of true predictions to total predictions. The f1-Score assesses the accuracy of theme classification, with recall indicating the sensitivity of the prediction, and precision reflecting the positive predictive value for each prediction [44]. To offer a comprehensive evaluation of classification accuracy, the f1-Score—a commonly used measure in text classification—strikes a balance between precision and recall. Hence, the f1-Score is the harmonic mean of recall and precision [45].

For each conceptual database, a partitioning approach was employed, allocating 70% of the annotated texts for training the algorithms and reserving the remaining 30% for testing purposes [46]. The objective was to establish a statistical likelihood for each algorithm, expressed through a prediction score, indicating the efficacy of classifying interactions. To adhere to recommended design practices, the training and testing sets were deliberately kept distinct for the calibration of each machine learning classifier [47]. Additionally, a tenfold cross-validation strategy was implemented for each algorithm, utilizing the K-Fold model from the Scikit-Learn module [37].

3. Results

3.1. Characteristics of the Participants

The interactions occurring in the verbatims of 35 patients were utilized by the five machine learning algorithms in this study for automated annotation. The details of the sampled patients can be found in Table 2.

Table 2. Characteristics of sampled participants.

Characteristics	Value (n = 18)
Sex (# male, # female)	16, 2
Age (mean in years)	42.6 ± 6.2
Education (mean in years)	12.8 ± 3.6
Ethnicity (Caucasian, others)	94.4%, 5.6%
% on Clozapine	61.1%

3.2. Performance of Ensemble Modeling

The performance accuracy of individual models and the ensemble model for automated classification of avatar interactions and patient interactions is found below.

3.2.1. Automated Classification of Avatar Interactions

The accuracy scores are presented in Figure 3 for all of the individual models on top of the ensemble model. The ensemble model performed the best, with a cross-validated accuracy of 0.71, closely followed by the LSVC at 0.66 and the MLP classifier at 0.66. The XGB performed with a cross-validated accuracy of 0.54, and the KNN algorithm performed with an accuracy of 0.57. The Multinomial NB performed the worst, with an accuracy of 0.48.

The mean metrics for recall, precision and f1-score for classification of the avatar interactions are presented in Table 3. It can be observed that the performances of all the metrics are consistent with the findings explicated for the accuracy, with the ensemble model achieving the best performance for accuracy, recall, precision and f1-score. This is closely followed by the LSVC and the MLP classifiers.

Table 3. Individual classifiers and ensemble mean scores for accuracy, precision, recall and f1-score for the classification of avatar interactions.

Models	Accuracy (Range)	Precision (Range)	Recall (Range)	f1-Score (Range)
LSVC	0.66 (0.64–0.67)	0.70 (0.69–0.71)	0.66 (0.65–0.67)	0.66 (0.65–0.67)
Multinomial NB	0.48 (0.47–0.48)	0.62 (0.47–0.49)	0.48 (0.47–0.49)	0.42 (0.41–0.43)
MLP	0.66 (0.64–0.67)	0.68 (0.65–0.69)	0.66 (0.65–0.67)	0.66 (0.65–0.67)
XGB	0.54 (0.54–0.55)	0.64 (0.64–0.65)	0.56 (0.56–0.57)	0.56 (0.56–0.57)
KNN	0.57 (0.55–0.58)	0.65 (0.63–0.67)	0.58 (0.56–0.60)	0.56 (0.54–0.58)
Ensemble	0.71 (0.69–0.72)	0.71 (0.69–0.72)	0.71 (0.69–0.72)	0.70 (0.69–0.71)

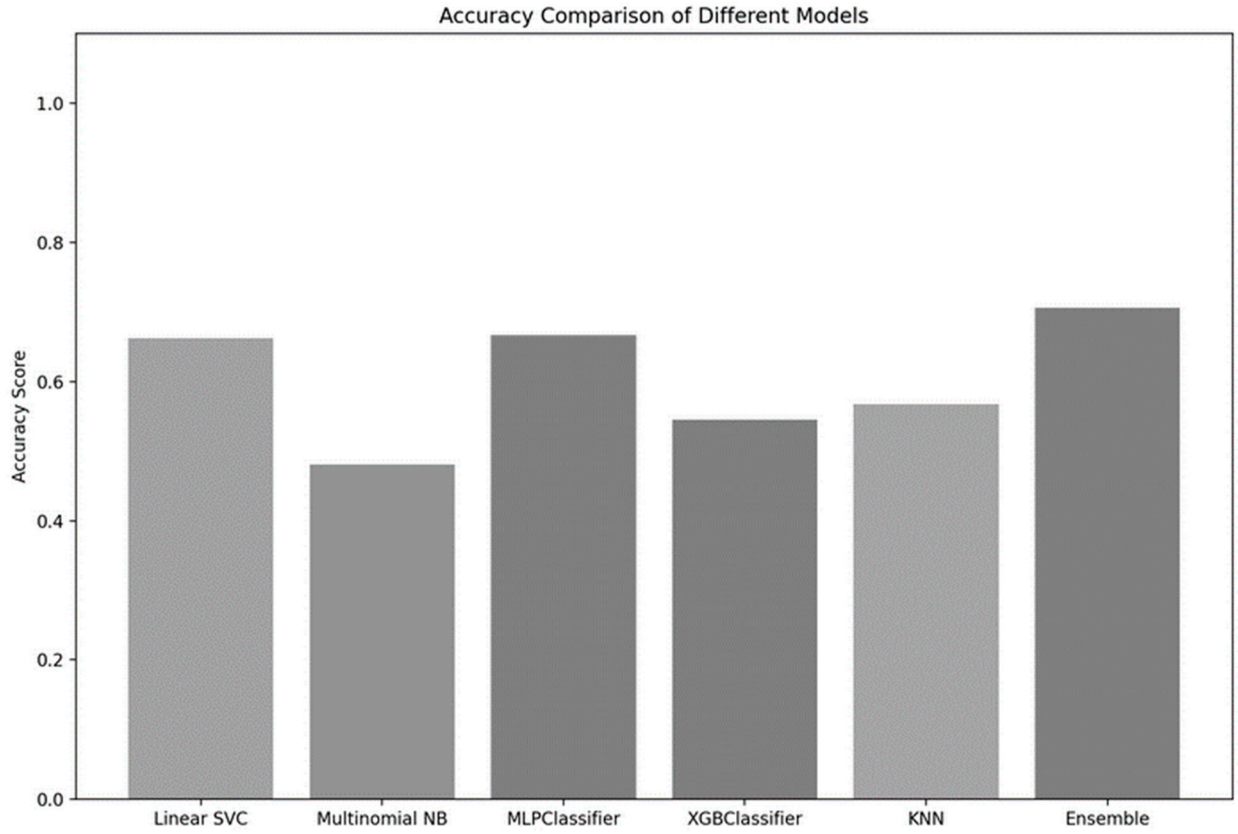

Figure 3. Accuracy comparison of the individual models implemented as well as the ensemble model which encompasses all the individual models' classification of avatar interactions.

3.2.2. Automated Classification of Patient Interactions

The accuracy scores are presented in Figure 4 for all the individual models coupled with the ensemble model. The ensemble model performed the best, with a cross-validated accuracy of 0.58. This is almost tied with the LSVC at 0.57 and the MLP classifier at 0.54. The XGB performed with a cross-validated accuracy of 0.48, and the KNN algorithm performed with an accuracy of 0.45. The Multinomial NB performed the worst, with an accuracy of 0.44.

The mean metrics for recall, precision and f1-score for classification of patient interactions are presented in Table 4. It can be observed that the performances of almost all the metrics are consistent with the findings explicated for the accuracy, with the ensemble model achieving the best performance for accuracy, recall, precision and f1-score. This is closely followed by the LSVC and the MLP classifiers. The only divergent metric found is that the LSVC performs better than the ensemble model for precision.

Table 4. Individual classifiers and ensemble mean scores for accuracy, recall, precision and f1-score for the classification of avatar interactions.

Models	Accuracy (Range)	Precision (Range)	Recall (Range)	f1-Score (Range)
LSVC	0.57 (0.56–0.58)	0.62 (0.60–0.63)	0.57 (0.56–0.58)	0.58 (0.57–0.9)
Multinomial NB	0.44 (0.44–0.45)	0.50 (0.50–0.51)	0.44 (0.43–0.44)	0.40 (0.39–0.41)
MLP	0.54 (0.53–0.55)	0.57 (0.55–0.57)	0.54 (0.53–0.55)	0.55 (0.54–0.56)
XGB	0.48 (0.48–0.49)	0.50 (0.49–0.51)	0.48 (0.48–0.49)	0.49 (0.48–0.50)
KNN	0.45 (0.43–0.46)	0.51 (0.48–0.53)	0.46 (0.45–0.47)	0.46 (0.45–0.47)
Ensemble	0.58 (0.57–0.9)	0.58 (0.57–0.9)	0.58 (0.57–0.9)	0.58 (0.57–0.9)

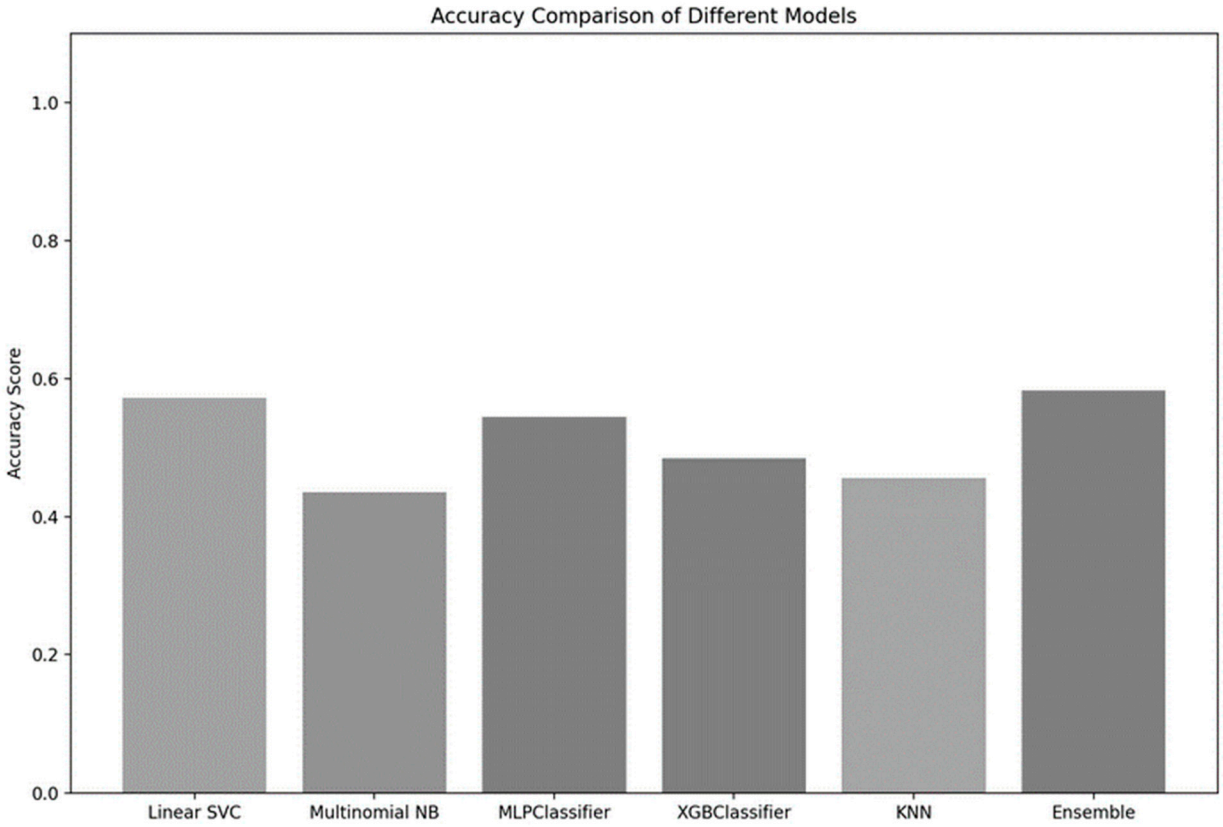

Figure 4. Accuracy comparison of the individual models implemented as well as the ensemble model which encompasses all of the individual models for the classification of patient interactions.

4. Discussion

The main objective of this study was to evaluate the change in accuracy of automated text classification machine learning algorithms when using an ensemble approach for immersive session verbatims of AT. For automatic classification of avatar and patient interactions, the ensemble approach performed the best in terms of classification accuracy. This was also the case for the recall, precision and f1-score metrics, apart from precision (in the classification of patient interactions), which was found to be better with the LSVC.

The performance of an ensemble model approach is often preferred over a single model when the data used are complex. This can hardly be compared to the literature in the context of psychotherapy considering that this has not been performed previously. However, as an example in the context of text classification, a recent study comparing the performances of several machine learning algorithms to an ensemble model comprising these algorithms, on a corpus comprising the Youtube Spam Collection Dataset and different text vectorization approaches, demonstrated that some of the ensemble (such as Adaboost and LightGBM) learning methods frequently produce enhanced text classification performance compared with base techniques [48]. It is also to be noted that the literature reviews on ensemble methods highlight that ensemble modeling is an acceptable technique for coping with individual classifiers' large variation while minimizing general mistakes [49]. Furthermore, ensemble techniques are reported to be an appropriate method to improve accuracy in text classification tasks, which is what has been observed in its use in AT [50]. Interestingly, a recent study comparing the use of single classifier to an ensemble approach in the domain of mental health suggests that for the prediction of mental health problems, ensemble models demonstrate better prediction results [51]. This could be similar for the appropriate prediction of patient interactions in the setting of psychotherapy as this is likewise established in textual instances.

The performance of LSVC for avatar and patient interactions was very similar to that of the ensemble model. LSVC was also found to perform better regarding precision in the patient interactions. This could be explained by the fact that most of the data were in fact linearly separable, as was previously assumed. Considering this sort of separability of the data, the data diversity decreased; therefore, the ensemble model compares to the best-performing linear classifier model, which is, in this scenario, the LSVC [52]. It could be hypothesized that the small amount of patients' data presented in the dataset also accounts for this observation, considering that, as more data become available, new themes could emerge from the verbatims and account for multicollinearity. This can be seen if two or more variables have linear correlations, which suggests that determining the marginal influence of a variable will be difficult [53].

The classification performances of the algorithms in the avatar conceptual dataset compared to the patient dataset indicated that interactions involving the avatar were classified with a higher overall accuracy. This can be explained by the fact that the classification complexity is reduced for the avatar as there are 13 possible themes for classification as compared to 14 for the patient interactions.

Potential future applications of ensemble modeling in the field of psychotherapy could achieve similar results as other ensemble modeling techniques in clinical psychiatry. For example, machine learning applications of ensemble models for the clinical information of 685 outpatients enabled the prediction of successful outcomes of cognitive behavioral therapy, with a balanced accuracy of 69% [54]. This sort of accuracy is comparable to that observed in our study. However, considering the limited number of studies that apply machine learning to psychotherapeutic content, it is clear at this stage that future studies are needed, notably on textual entities such as therapeutic interactions.

It is also important to note practical ethical considerations when using such techniques for psychotherapeutic interventions. In this study, considering that the data were processed by several machine learning algorithms, they were anonymized to ensure the confidentiality and privacy of the patients. The accountability of data being automatically categorized is also the responsibility of the clinician when machine learning is applied to a clinical context, and this should be further investigated [55].

Limitations

The models utilized in constructing the ensemble model are currently limited by the relatively small databases available for Avatar Therapy (AT). The performance trend of the ensemble model will be re-evaluated as more patients are added to the dataset. It is important to note that the transcripts analyzed in this study were written in Canadian French, and obtaining vectorizers that included stop words specific to Canadian French proved challenging. Stop words, which are often excluded during tokenization due to their limited meaning, may impact the analysis' accuracy. The lack of sufficient stop words in Canadian French could result in the inclusion of inconsequential terms. Regarding the patient conceptual database, it is noteworthy that three-fifths of the individual algorithms initially achieved a classification accuracy below 0.5, limiting the performance of the ensemble models.

5. Conclusions

In conclusion, this study evaluated the change in accuracy of automated text classification machine learning algorithms when using an ensemble approach in immersive session verbatims of AT. Automated classification of text is not a simple task when considering psychotherapeutic interventions, and this study demonstrated that ensemble modeling performed best in terms of the accuracy of the classification of avatar and patient interactions. This technique also performed better than its individual counterparts for precision, recall and f1-score. The only exception was the precision of the classification of patient interactions, for which the LSVC performed best. This study offers the first evaluation of ensemble modeling in the context of AT and provides an objective optimized approach in

the classification of textual interactions based on immersive session verbatims. This technique might be used in future research to give insight into the interactions being classified and the therapeutical response of patients based on their experience with AT immersion sessions with optimized precision by employing an ensemble methodological approach.

Author Contributions: Conceptualization, A.H., K.P., S.P. and A.D.; methodology, A.H. and A.D.; validation, A.H. and A.D.; formal analysis, A.H.; investigation, A.H.; data curation, A.H.; writing—original draft preparation, A.H.; writing—review and editing, A.H., K.P, S.P. and A.D.; supervision, K.P., S.P. and A.D.; project administration, K.P.; funding acquisition, K.P., S.P. and A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was indirectly supported by Le Fonds de recherche du Québec—Santé (FRQS); Otsuka Canada Pharmaceutical Inc.; Chaire Eli Lilly Canada de recherche en schizophrénie; MEI (Ministère de l'Économie et de l'Innovation); Services et recherches psychiatriques AD; Fonds d'excellence en recherche Apogée Canada. These fundings bodies had no part in the data collection, analysis, interpretation of data and in writing the manuscript.

Institutional Review Board Statement: This study was approved by the institutional ethical committee, and written informed consent was obtained from all patients. Patients that are part of this study were selected based on the proof-of-concept trial from Percy du Sert 2018's study and Dellazizzo 2021's study [15]. The trial was conducted in accordance with the Declaration of Helsinki and was approved by the institutional ethical committee (CER IPPM 16-17-06). We obtained written informed consent from all patients.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets generated and/or analyzed during the current study are not publicly available due to patients' privacy but are available from the corresponding author on reasonable request.

Acknowledgments: "Database icon" by Nurul Hotimah is licensed under Creative Commons BY 3.0 https://creativecommons.org/licenses/by/3.0/ (accessed on 11 August 2023).

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

References

1. Carrà, G.; Crocamo, C.; Angermeyer, M.; Brugha, T.; Toumi, M.; Bebbington, P. Positive and negative symptoms in schizophrenia: A longitudinal analysis using latent variable structural equation modelling. *Schizophr. Res.* **2018**, *204*, 58–64. [CrossRef] [PubMed]
2. Correll, C.U.; Schooler, N.R. Negative Symptoms in Schizophrenia: A Review and Clinical Guide for Recognition, Assessment, and Treatment. *Neuropsychiatr. Dis. Treat.* **2020**, *16*, 519–534. [CrossRef]
3. Patel, K.R.; Cherian, J.; Gohil, K.; Atkinson, D. Schizophrenia: Overview and treatment options. *Peer Rev. J. Formul. Manag.* **2014**, *39*, 638–645.
4. Stępnicki, P.; Kondej, M.; Kaczor, A.A. Current Concepts and Treatments of Schizophrenia. *Molecules* **2018**, *23*, 2087. [CrossRef]
5. Siskind, D.; Siskind, V.; Kisely, S. Clozapine Response Rates among People with Treatment-Resistant Schizophrenia: Data from a Systematic Review and Meta-Analysis. *Can. J. Psychiatry* **2017**, *62*, 772–777. [CrossRef]
6. Corripio, I.; Roldán, A.; Sarró, S.; McKenna, P.J.; Alonso-Solís, A.; Rabella, M.; Díaz, A.; Puigdemont, D.; Pérez-Solà, V.; Álvarez, E.; et al. Deep brain stimulation in treatment resistant schizophrenia: A pilot randomized cross-over clinical trial. *EBioMedicine* **2020**, *51*, 102568. [CrossRef]
7. Huckle, P.L.; Palia, S.S. Managing resistant schizophrenia. *Br. J. Hosp. Med.* **1993**, *50*, 467–471. [PubMed]
8. Suzuki, T.; Remington, G.; Mulsant, B.H.; Uchida, H.; Rajji, T.K.; Graff-Guerrero, A.; Mimura, M.; Mamo, D.C. Defining treatment-resistant schizophrenia and response to antipsychotics: A review and recommendation. *Psychiatry Res.* **2012**, *197*, 1–6. [CrossRef] [PubMed]
9. Correll, C.U.; Howes, O.D. Treatment-Resistant Schizophrenia: Definition, Predictors, and Therapy Options. *J. Clin. Psychiatry* **2021**, *82*. [CrossRef]
10. Chakrabarti, S. Clozapine resistant schizophrenia: Newer avenues of management. *World J. Psychiatry* **2021**, *11*, 429–448. [CrossRef]
11. Potkin, S.G.; Kane, J.M.; Correll, C.U.; Lindenmayer, J.-P.; Agid, O.; Marder, S.R.; Olfson, M.; Howes, O.D. The neurobiology of treatment-resistant schizophrenia: Paths to antipsychotic resistance and a roadmap for future research. *Schizophrenia* **2020**, *6*, 1. [CrossRef]

12. Bighelli, I.; Huhn, M.; Schneider-Thoma, J.; Krause, M.; Reitmeir, C.; Wallis, S.; Schwermann, F.; Pitschel-Walz, G.; Barbui, C.; Furukawa, T.A.; et al. Response rates in patients with schizophrenia and positive symptoms receiving cognitive behavioural therapy: A systematic review and single-group meta-analysis. *BMC Psychiatry* **2018**, *18*, 380. [CrossRef]
13. Morrison, A.P.; Pyle, M.; Gumley, A.; Schwannauer, M.; Turkington, D.; MacLennan, G.; Norrie, J.; Hudson, J.; E Bowe, S.; French, P.; et al. Cognitive behavioural therapy in clozapine-resistant schizophrenia (FOCUS): An assessor-blinded, randomised controlled trial. *Lancet Psychiatry* **2018**, *5*, 633–643. [CrossRef] [PubMed]
14. Aali, G.; Kariotis, T.; Shokraneh, F. Avatar Therapy for people with schizophrenia or related disorders. *Cochrane Database Syst. Rev.* **2020**, *2020*, Cd011898. [CrossRef]
15. Leff, J.; Williams, G.; Huckvale, M.A.; Arbuthnot, M.; Leff, A.P. Computer-assisted therapy for medication-resistant auditory hal-lucinations: Proof-of-concept study. *Br. J. Psychiatry* **2013**, *202*, 428–433. [CrossRef] [PubMed]
16. Bisso, E.; Signorelli, M.S.; Milazzo, M.; Maglia, M.; Polosa, R.; Aguglia, E.; Caponnetto, P. Immersive Virtual Reality Applications in Schizo-phrenia Spectrum Therapy: A Systematic Review. *Int. J. Env. Res. Public. Health* **2020**, *17*, 6111. [CrossRef]
17. Leff, J.; Williams, G.; Huckvale, M.; Arbuthnot, M.; Leff, A.P. Avatar therapy for persecutory auditory hallucinations: What is it and how does it work? *Psychosis* **2014**, *6*, 166–176. [CrossRef] [PubMed]
18. Craig, T.K.; Rus-Calafell, M.; Ward, T.; Leff, J.P.; Huckvale, M.; Howarth, E.; Emsley, R.; Garety, P.A. AVATAR therapy for auditory verbal hallucinations in people with psychosis: A single-blind, randomised controlled trial. *Lancet Psychiatry* **2017**, *5*, 31–40. [CrossRef]
19. Dellazizzo, L.; Potvin, S.; Phraxayavong, K.; Dumais, A. One-year randomized trial comparing virtual reality-assisted therapy to cognitive–behavioral therapy for patients with treatment-resistant schizophrenia. *Schizophrenia* **2021**, *7*, 9. [CrossRef]
20. Dellazizzo, L.; du Sert, O.P.; Phraxayavong, K.; Potvin, S.; O'Connor, K.; Dumais, A. Exploration of the dialogue components in Avatar Therapy for schizophrenia patients with refractory auditory hallucinations: A content analysis. *Clin. Psychol. Psychother.* **2018**, *25*, 878–885. [CrossRef]
21. Beaudoin, M.; Potvin, S.; Machalani, A.; Dellazizzo, L.; Bourguignon, L.; Phraxayavong, K.; Dumais, A. The therapeutic processes of avatar therapy: A content analysis of the dialogue between treatment-resistant patients with schizophrenia and their avatar. *Clin. Psychol. Psychother.* **2021**, *28*, 500–518. [CrossRef] [PubMed]
22. Szymańska, A.; Dobrenko, K.; Grzesiuk, L. Characteristics and experience of the patient in psychotherapy and the psychotherapy's effectiveness. *A Struct. Approach. Psychiatr. Pol.* **2017**, *51*, 619–631. [CrossRef]
23. Runciman, W.B. Qualitative versus quantitative research—Balancing cost, yield and feasibility. *Qual. Saf. Health Care* **2002**, *11*, 146–147. [CrossRef] [PubMed]
24. Pannucci, C.J.; Wilkins, E.G. Identifying and Avoiding Bias in Research. *Plast. Reconstr. Surg.* **2010**, *126*, 619–625. [CrossRef] [PubMed]
25. Althubaiti, A. Information bias in health research: Definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* **2016**, *9*, 211–217. [CrossRef] [PubMed]
26. Dogra, V.; Verma, S.; Kavita; Chatterjee, P.; Shafi, J.; Choi, J.; Ijaz, M.F. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Comput. Intell. Neurosci.* **2022**, *2022*, 1883698. [CrossRef]
27. Jovel, J.; Greiner, R. An Introduction to Machine Learning Approaches for Biomedical Research. *Front. Med.* **2021**, *8*, 771607. [CrossRef]
28. Hey, T.; Butler, K.; Jackson, S.; Thiyagalingam, J. Machine learning and big scientific data. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2020**, *378*, 20190054. [CrossRef]
29. Hudon, A.; Beaudoin, M.; Phraxayavong, K.; Dellazizzo, L.; Potvin, S.; Dumais, A. Use of Automated Thematic Annotations for Small Data Sets in a Psychotherapeutic Context: Systematic Review of Machine Learning Algorithms. *JMIR Ment. Health* **2021**, *8*, e22651. [CrossRef]
30. Hudon, A.; Beaudoin, M.; Phraxayavong, K.; Dellazizzo, L.; Potvin, S.; Dumais, A. Implementation of a machine learning algorithm for automated thematic annotations in avatar: A linear support vector classifier approach. *Health Inform. J.* **2022**, *28*, 14604582221142442. [CrossRef] [PubMed]
31. Hudon, A.; Couture, J.; Dellazizzo, L.; Beaudoin, M.; Phraxayavong, K.; Potvin, S.; Dumais, A. Dyadic Interactions of Treatment-Resistant Schizophrenia Patients Having Followed Virtual Reality Therapy: A Content Analysis. *J. Clin. Med.* **2023**, *12*, 2299. [CrossRef]
32. Bhavsar, H.; Ganatra, A. A comparative study of training algorithms for supervised machine learning. *Int. J. Soft Comput. Eng. (IJSCE)* **2012**, *2*, 2231–2307.
33. Ganaie, M.A.; Minghui, H.; Malik, A.K.; Tanveer, M.; Suganthan, P.N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
34. Verma, A.; Mehta, S. A comparative study of ensemble learning methods for classification in bioinformatics. In Proceedings of the 2017 7th International Conference on Cloud Computing, Data Science & Engineering—Confluence (Confluence), Noida, India, 12–13 January 2017; pp. 155–158.
35. Pintelas, P.; Livieris, I.E. Special Issue on Ensemble Learning and Applications. *Algorithms* **2020**, *13*, 140. [CrossRef]
36. Lewis, R.B.; Maas, S.M. QDA Miner 2.0: Mixed-model qualitative data analysis software. *Field Methods* **2007**, *19*, 87–108. [CrossRef]
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

38. Hudon, A.; Beaudoin, M.; Phraxayavong, K.; Potvin, S.; Dumais, A. Unsupervised Machine Learning Driven Analysis of Verbatims of Treatment-Resistant Schizophrenia Patients Having Followed Avatar Therapy. *J. Pers. Med.* **2023**, *13*, 801. [CrossRef] [PubMed]
39. Chen, J.; Yuan, P.; Zhou, X.; Tang, X. (Eds.) *Performance Comparison of TF*IDF, LDA and Paragraph Vector for Document Classification*; Springer: Singapore, 2016.
40. Kabari, L.G.; Onwuka, U.C. Comparison of bagging and voting ensemble machine learning algorithm as a classifier. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2019**, *9*, 19–23.
41. Peppes, N.; Daskalakis, E.; Alexakis, T.; Adamopoulou, E.; Demestichas, K. Performance of machine learning-based multi-model voting ensemble methods for network threat detection in agriculture 4.0. *Sensors* **2021**, *21*, 7475. [CrossRef] [PubMed]
42. Alsulami, B.; Almalawi, A.; Fahad, A. Toward an Efficient Automatic Self-Augmentation Labeling Tool for Intrusion Detection Based on a Semi-Supervised Approach. *Appl. Sci.* **2022**, *12*, 7189. [CrossRef]
43. Manconi, A.; Armano, G.; Gnocchi, M.; Milanesi, L. A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19. *Appl. Sci.* **2022**, *12*, 7554. [CrossRef]
44. Goutte, C.; Gaussier, E. (Eds.) A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2005.
45. Opitz, J.; Burst, S. Macro f1 and macro f1. *arXiv* **2019**, arXiv:191103347.
46. Gholamy, A.; Kreinovich, V.; Kosheleva, O. *Why 70/30 or 80/20 Relation between Training and Testing Sets: A Pedagogical Explanation*; The University of Texas at El Paso: El Paso, TX, USA, 2018.
47. Birba, D.E. *A Comparative Study of Data Splitting Algorithms for Machine Learning Model Selection*; Kth Royal Institute of Technology: Stockholm, Sweden, 2020; Available online: https://www.diva-portal.org/smash/get/diva2:1506870/FULLTEXT01.pdf (accessed on 2 July 2023).
48. Ibrahim, Y.; Okafor, E.; Yahaya, B.; Yusuf, S.M.; Abubakar, Z.M.; Bagaye, U.Y. Comparative Study of Ensemble Learning Techniques for Text Classification. In Proceedings of the 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, 15–16 July 2021; pp. 1–5.
49. Ammar, M.; Rania, K. An effective ensemble deep learning framework for text classification. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34 Pt A*, 8825–8837.
50. Palanivinayagam, A.; El-Bayeh, C.Z.; Damaševičius, R. Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. *Algorithms* **2023**, *16*, 236. [CrossRef]
51. Chung, J.; Teo, J. Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain Inform.* **2023**, *10*, 1. [CrossRef]
52. Dietterich, T.G. (Ed.) Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000.
53. Chan, J.Y.-L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.-W.; Chen, Y.-L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics* **2022**, *10*, 1283. [CrossRef]
54. Taubitz, F.-S.; Büdenbender, B.; Alpers, G.W. What the future holds: Machine learning to predict success in psychotherapy. *Behav. Res. Ther.* **2022**, *156*, 104116. [CrossRef] [PubMed]
55. Habli, I.; Lawton, T.; Porter, Z. Artificial intelligence in health care: Accountability and safety. *Bull. World Health Organ.* **2020**, *98*, 251–256. [CrossRef] [PubMed]