Received: 15 March 2024 | Revised: 15 May 2024 | Accepted: 29 May 2024 | Published online: 5 June 2024

RESEARCH ARTICLE

Medinformatics
yyyy, Vol. XX(XX) 1–5
DOI: 10.47852/bonviewMEDIN42022849

Diagnosis of Schizophrenia and Psychosis in Māori People Using Speech Assay: A Natural Language Processing Approach

Randall Ratana[1,] *, Hamid Sharifzadeh[2] and Jamuna G Krishnan[3]

1 Unitec Institute of Technology, New Zealand, randall@randallratana.nz, ORCID: https://orcid.org/0000-0001-6154-021X
2 Unitec Institute of Technology, New Zealand, hamid.sharifzadeh@unitec.ac.nz
3 Royal Darwin Hospital, Australia, jamuna.krishnan@nt.gov.au

***Corresponding author:** Randall Ratana, Unitec Institute of Technology, New Zealand, randall@randallratana.nz

Abstract: Indigenous peoples and ethnic minorities worldwide have a higher incidence of psychosis, primarily schizophrenia. In New Zealand, Māori individuals experience greater rates of anxiety and depression compared to non-Māori, with a significantly higher incidence of mental illness among Māori. We propose natural language processing (NLP) trained on speech samples from Māori patients as a potential solution to the problem of culturally biased psychometric screening tools for psychosis and schizophrenia. This research examines NLP's ability to diagnose psychosis in Māori patients by analysing speech and language abnormalities as indicators of severe mental illnesses such as schizophrenia. Our research emphasizes the need for inclusive language models and investigates cross-cultural applicability. We employed a three-part method: conducting clinical interviews, pre-processing data with the Natural Language Toolkit (NLTK) and applying language classifiers. The study's results demonstrate the promise of NLP, but limited patient data necessitates further research, including standardizing datasets and integrating NLP with indigenous languages. This research represents a step towards improving diagnostic accuracy and support for Māori people suffering from psychosis, aligning with healthcare's goal of fair and culturally responsive mental health screening.

Keywords: psychosis, schizophrenia, Māori healthcare, natural language processing (NLP), machine learning

1. Introduction

Indigenous individuals and ethnic minorities have greater rates of schizophrenia and psychosis than other ethnicities [1]. In New Zealand (NZ), Māori individuals experience greater rates of anxiety and depression compared to non-Māori individuals, with mental illness rates higher among Māori males than females [2]. According to 2008-2015 mental health statistics, Māori males had a higher 12-month prevalence of schizophrenia compared to other ethnicities, including European New Zealanders [3-5].

In 2019, the NZ Mental Health Inquiry found that Māori individuals suffered mental anguish at a rate 50% greater than did non-Māori [6]. Furthermore, Māori individuals were found to be 30% more likely to remain undiagnosed for mental illness. The inquiry revealed systemic issues in diagnosing Māori individuals correctly, leading to low outcomes across indicators and diagnoses. A significant issue is the absence of a culturally appropriate mental health assessment approach that considers indigenous worldviews on mental illness [7, 8].

Existing assessment techniques for psychotic disorders are known to contain cultural biases that do not accurately address the mental health needs of the Māori population [9, 10]. Moreover, the current diagnostic methods rely on the universal assumption that schizophrenia presents uniformly across cultures [11, 12]. However, indigenous experiences of schizophrenia and psychosis differ in content and meaning. Natural language processing (NLP) offers a more nuanced approach to identifying sub-linguistic distinctions in mental illness [11, 13-15].

This research aims to diagnose psychosis in Māori patients using natural language analysis to identify speech and language dysfunction as markers of mental illness. The goal is to assist psychiatrists and frontline mental health physicians in assessing Māori individuals with psychotic symptoms. We employed a three-part method that involved clinical interviews, pre-processing data using the Natural Language Toolkit (NLTK) and applying language classifiers to identify relevant linguistic features [16, 17].

By leveraging NLP, we aim to create inclusive language models and investigate the cross-cultural applicability of this approach. The research acknowledges the need for culturally responsive mental health assessments while demonstrating how analyzing speech patterns can reveal important insights into schizophrenia diagnosis for Māori patients.

2. Materials and Methods

2.1. Study protocol

2.1.1. Clinical interviews

To develop an early psychosis-screening tool for young Māori, we collected speech samples from individuals at risk of psychosis and those diagnosed with schizophrenia. The clinical interview used in this research was a method of eliciting speech. In the absence of a prompt-based protocol, this research captured the patient's narrative during the clinical interview, which consisted of a narrative inquiry approach by the interviewer. The average length of the speech sample was set at 10 minutes in both the patient sample and control sample. We employed the questions outlined in Table 1 to engage with the community about this project.

Table 1. Objectives of the research for Māori community

Research Objective / Purpose	Statement
OB-1	What research do we want to conduct?
OB-2	Whom is that research for?
OB-3	What difference will it make?
OB-4	Who will carry out this research?
OB-5	How do we want the research to be done?
OB-6	Who will benefit?

After verifying the hypothesis, we applied predictive machine learning models to the text in step three.

2.2. Speech assay analysis

Speech data was analyzed using Latent Semantic Analysis (LSA). While most studies have focused on either statistical language analysis or acoustic analysis, this study focused on language analysis only. The study was conducted at a single site, and cross-site validation was not considered.

2.3. Participants

2.3.1. Patient group

We identified patients who were referred to clinics and considered at risk of psychosis or who had experienced their first episode of psychosis or schizophrenia. Additionally, patients discharged from the inpatient unit after a psychotic illness episode was included. The Brief Psychiatric Rating Scale (BPRS) was used to measure baseline symptoms of psychosis based on a clinical interview, while CAARMS helped indicate a history of attenuated positive psychotic symptoms.

2.3.2. Control group

The control group consisted of six individuals randomly selected from close-knit family and friends who were aware of the research project. Their ages ranged from 40 to 70 years, and all identified as Māori. They were given 10 minutes to talk about themselves and asked the same questions posed to the rangatahi Māori participants. All participants disclosed no known past psychiatric history.

2.3.3. Sample size and sampling procedure

The sample size was based on a pilot study and used purposive sampling to identify Māori Rangatahi who was referred by primary care providers or other services to secondary mental health services for assessment due to abnormal mental state. However, the high-needs population and complex psychiatric histories, including poly-substance abuse and intellectual disabilities, meant strict inclusion and exclusion criteria limited participant recruitment. Ultimately, only eight out of fifteen participants who met the study criteria consented to participate. Table 2 details participant characteristics from this locality.

Table 2 Participant demographics

Characteristic	Criteria
Age	18–40
Ethnicity	NZ Māori
Gender	Female or male

2.4 Computational methods

2.4.1. Preliminary phase of speech analysis

The methods employed included the implementation of speech elicitation techniques (free-talk approach using a series of questions), transcription of audio-recorded data (clinical interview), and data preparation as an essential initial step. The Natural Language Toolkit (NLTK) was used to rigorously clean text data of noisy and non-linguistic components [18]. POS tagging, machine learning algorithm classification, and voice analysis were used to compare patient and control data's language. Word cloud visual representations were examined to identify commonly recurring terms in text. Research data was also analyzed using feature comparison.

2.4.2. Generation of synthetic data

Due to the limited size of the patient dataset, base data was segmented. Segmenting continuous speech samples into meaningful parts was necessary to study speech patterns, sentiment, and psychosis-related linguistic markers. Words such as "hear," "voice," "sound," "loud," "chant," and "whisper" helped segment data for voice recognition (auditory hallucination) and understanding semantic meanings.

Synthetic data was generated using OpenAI's GPT-3.5 Turbo model, employing a random sampling method to select diverse synthetic data regions for comparison with real dataset portions. Preliminary analysis of the original dataset was conducted using Latent Dirichlet Allocation (LDA) to identify patterns and calculate average word embeddings. The LDA chart revealed patterns that underscored the preservation of language attributes.

2.4.3. Second phase of speech analysis

The data for this study was derived from interviews with hospitalized patients who met specific criteria (Table 3). Initial steps included transcription and dataset cleaning using POS tagging and NLTK for word lemmatization, leading to the creation of a patient-specific word2vec model. This model clusters commonly used words by frequency (TF-IDF) and context.

Table 3. Interview questionnaire

Question No	Statement
Q-1	What is your Name?
Q-2	What is your Profession?
Q-3	Can you tell me about yourself?
Q-4	Why are you in the hospital?
Q-5	What are your symptoms?
Q-6	How have you found it in the hospital?
Q-7	What are your future goals?

The processed text from both patient and control groups was compared using word cloud and Latent Semantic Analysis (LSA) on TF-IDF vectors, revealing distinct word usage patterns between the groups. Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) helped identify textual similarities and differences by reducing data dimensionality. Latent Dirichlet Allocation (LDA) was used to ascertain relevant themes in the group texts and to validate the synthetic data.

Figures 1A and 1B illustrate these differences, showing a broader but less semantically associated range of words used by patients compared to controls. Figures 2A and 2B display the frequency of word usage, indicating that the patient group used a more limited and repetitive vocabulary, often lacking in organization and demonstrating a tendency to add new elements to their answers.

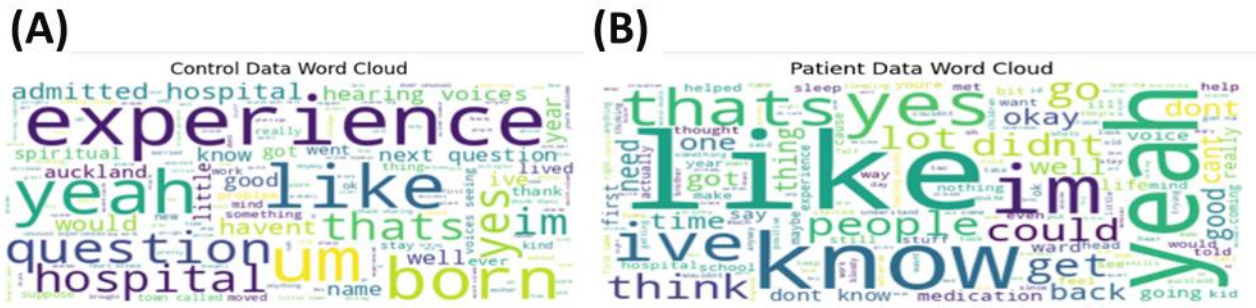

Figure 1 (A) Control data word cloud. (B) Patient data word cloud

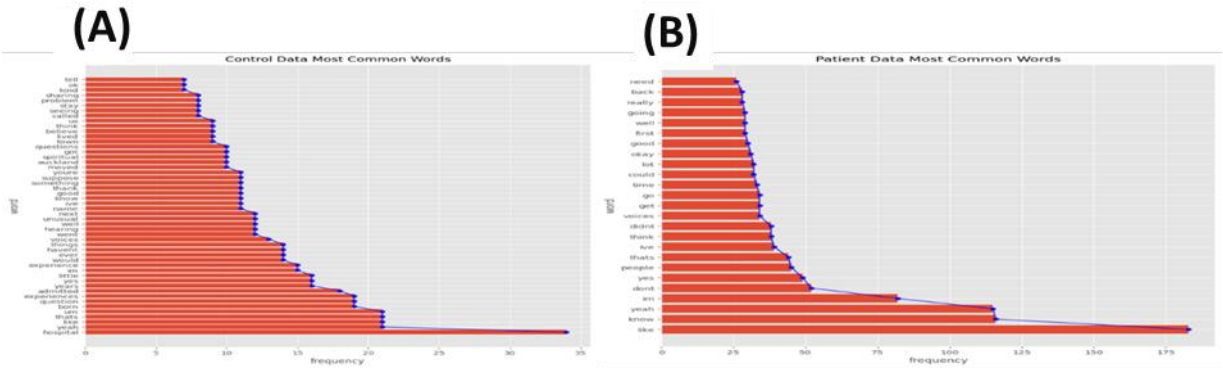

Figure 2 (A) Frequency chart of words in control group. (B) Frequency chart of words in patient group

2.4.5. Third phase of speech analysis

Semantic density and coherence were calculated using a method called multi-unpacking, utilizing the NLTK library to preprocess and transcribe speech. After transcription, nouns, pronouns, and verbs were selected from the lemmatized phrases. The Gensim Word2Vec model calculated vector forms of these words along with context words and adverbs, based on the Reddit pre-trained model. The model identifies relationships between words and represents their proximity within the text.

Patients in the study frequently used terms like "living," "lives," "lived," and "moved." The Word2Vec model leverages this data to enhance understanding of word relationships. Additional processing includes removing stop words and lemmatizing words to their base forms (e.g., reducing "suffering" to "suffer").

The Reddit corpus and Word2Vec model helped extract contextual terms for each phrase. Semantic density was then calculated for each statement, and the overall semantic density of the patient interviews was compared to the control group. Regression analysis was performed to assess the models' accuracy, testing logistic regression, support vector machine, random forest, and neural network models [19].

2.5. Ethical approval and data protection

Due to the Māori culture focus, the study required ethical approval from both the local Māori ethics committee and the Health and Disability Ethics Committee (HDEC). Consultation with a local Māori research organization was essential, and a recognized kaumātua (elder) provided cultural guidance [20]. Data storage was managed by a co-investigator from the Bay of Plenty District Health Board, with encrypted data kept private. Participants were encouraged to involve their whānau (family), who could attend talks, redact sensitive information, and approve recordings [21].

3. Results and Discussion

3.1. Preliminary speech analysis findings

3.1.1. Word usage patterns using word cloud

The study employed was elicitation techniques using a free-talk approach, where the interview data were transcribed, and noise and non-linguistic components were removed via the Natural Language Toolkit (NLTK). Word Clouds visually represented recurring terms in the text to identify differences in word usage between patient and control groups. This analysis revealed notable differences in the vocabulary between the two groups, with words like "experience" and "hospital" being more common among the control group, while patients frequently used terms such as "yeah" and "like".

3.1.2. Differences highlighted through Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA)

To distinguish linguistic traits between patients and controls, the study applied Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). LSA produced a 2D matrix visualized through heatmaps to show the relationship between control and patient files. The analysis showed higher similarity within the same groups (Control-Control or Patient-Patient) than between different groups. Figures 3 and 4 illustrate this comparison. Furthermore, Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) were employed to visualize the separation between the two groups.

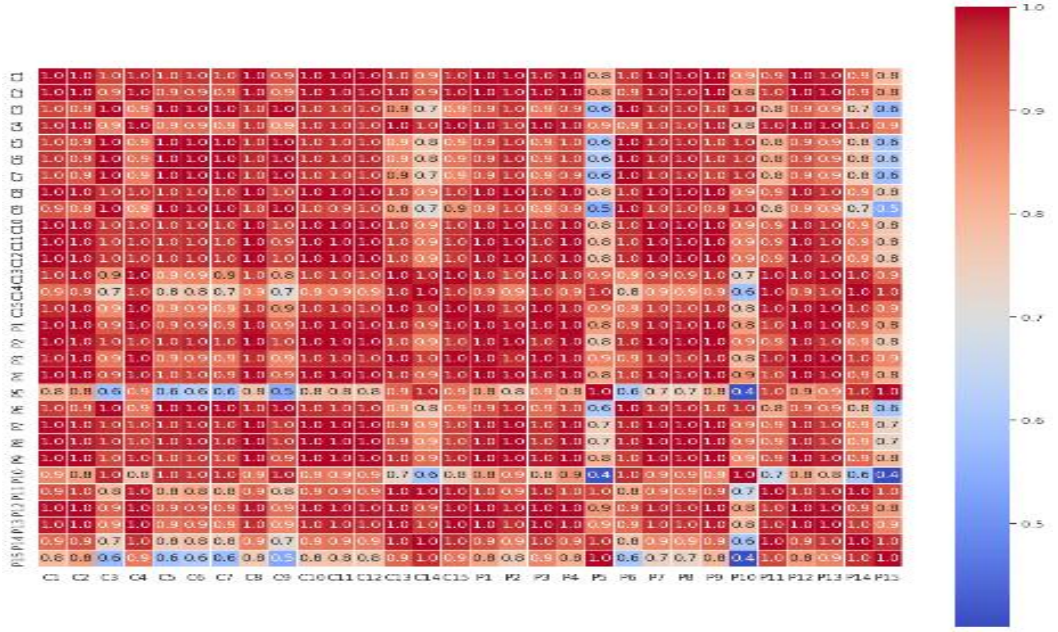

Figure 3.Heat map of LSA matrix

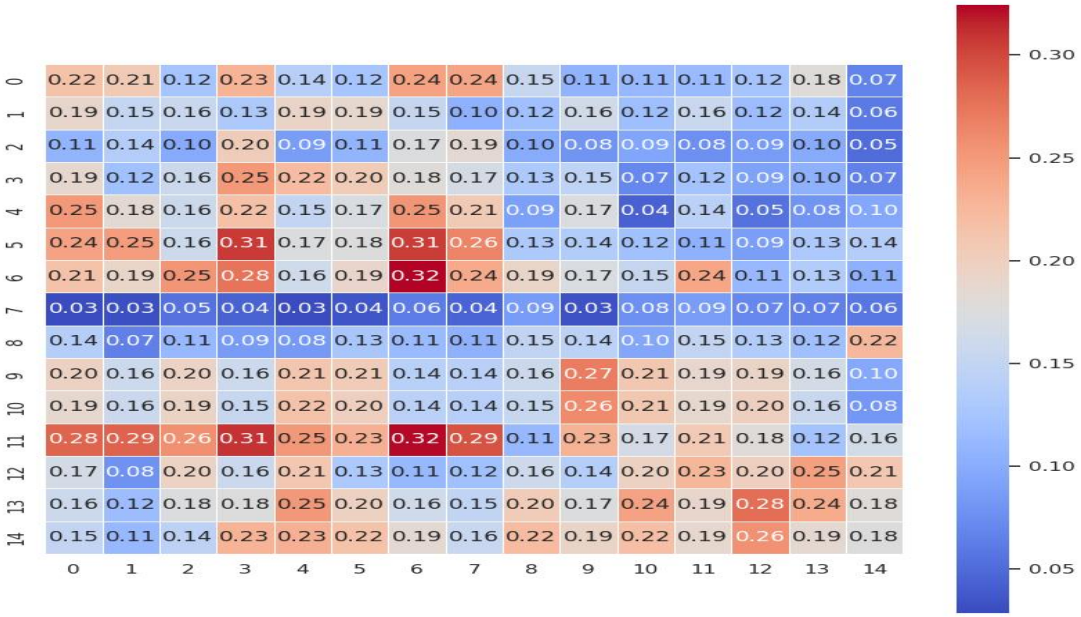

Figure 4. Similarity chart between control and patient data

Figure 5 shows the scatter plot, revealing clear differences between control and patient data.

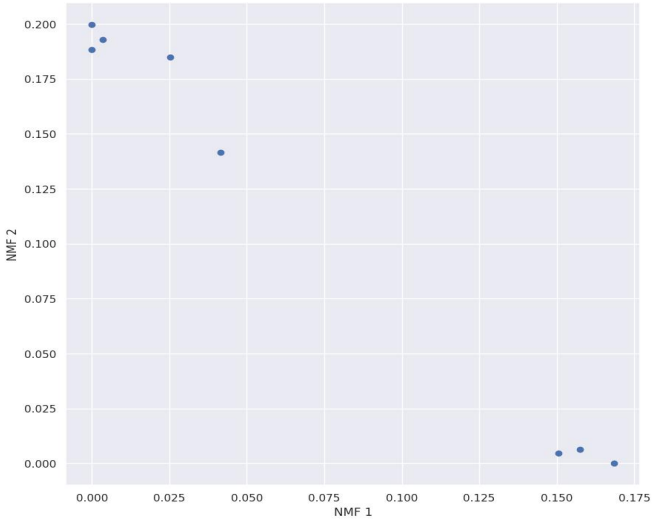

Figure 5. NMF scatter plot

3.1.3. Heatmaps and scatter plots illustrating variations in linguistic features

Figure 3's heatmap shows that the similarity between control and patient files is generally lower compared to within-group comparisons. The darker shades indicate higher similarity, while the lighter shades signify lower similarity. In the NMF scatter plot in Figure 5, two separate clusters highlight differences between patient and control groups.

3.2. Generation of synthetic data

3.2.1. Approach used to generate synthetic data using GPT-3.5 Turbo

To counter the limited size of the patient dataset, the base data were segmented, with segments used to generate synthetic data via Open AI's GPT-3.5 Turbo model. Text segments were used to preserve language accuracy and consistency while maintaining adherence to the original segments.

3.2.2. Comparative analysis between real and synthetic datasets using LDA

After adding synthetic data to the real dataset, the LDA model calculated mean word embeddings for each data class, demonstrating how synthetic data affected the precision of patient information. Figures 6A and 6B show that patterns observed before incorporating synthetic data closely resemble those after its inclusion.

3.2.3. Validation through matching patterns before and after incorporating synthetic data

Comparing Figure 6A (before synthetic data) and Figure 6B (after synthetic data), the analysis confirmed that synthetic data generation did not compromise the dataset's integrity. The synthetic data complemented the original dataset, improving the confidence in patient data classification.

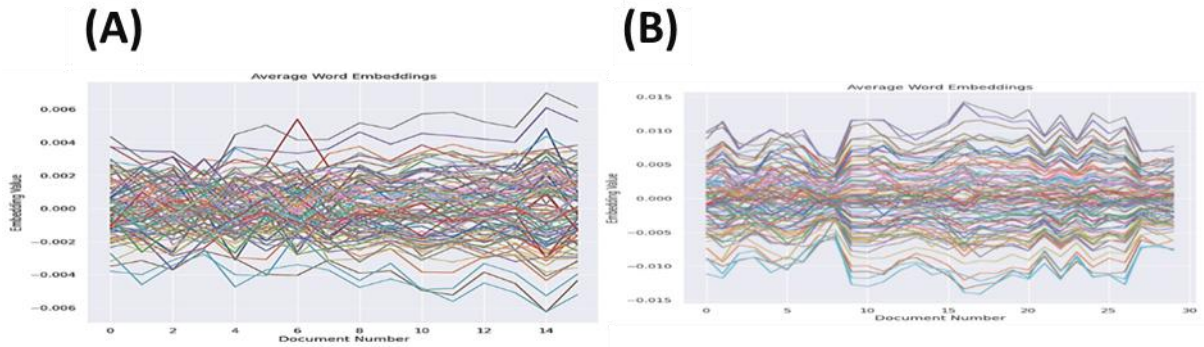

Figure 6 (A) Average word embedding of LDA before synthetic data. (B) Average word embedding of LDA after synthetic data

The analysis established that natural language processing techniques could effectively differentiate linguistic patterns between Māori patients with schizophrenia and the control group. Despite challenges with sample size and cultural nuances, these preliminary results confirmed the potential of machine learning models in assisting early psychosis detection in indigenous communities.

3.3. Advanced speech analysis techniques

3.3.1. Construction of a patient-specific Word2Vec model for semantic clustering

In the second phase of speech analysis, the transcribed data from patient interviews were cleaned and processed using NLTK for lemmatization and Part of Speech (POS) tagging. A patient-specific Word2Vec model was constructed for clustering commonly used words according to frequency and context. This allowed terms with similar contextual usage, such as 'mania,' 'depression,' and 'mood disorder,' to be grouped together. The Word2Vec model facilitated semantic clustering and identified relationships between linguistic terms used by the patients.

3.3.2. Feature comparison using Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF)

Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) were used to visualize the distinction between control and patient data. The scatter plot (Figure 5) shows clear clustering, with control and patient groups being distinguishable. This visualization highlights the distinct patterns between the two groups, confirming the differences in semantic features. PCA reduced data dimensionality, while NMF emphasized the variance between control and patient data.

3.3.3. Identification of topic correlations in speech samples via Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA), a probabilistic model, identified topics across speech samples. It assumes that each document is a mix of various themes, and each word corresponds to a specific topic. LDA training on patient and control corpora revealed topic correlations and frequently occurring words within each group, thereby enabling a better understanding of linguistic behavior. Figures 6A and 6B show the comparison of average word embeddings before and after adding synthetic data, confirming that data integrity was maintained.

3.4. Predictive model results

3.4.1. Evaluation of predictive models including logistic regression, random forest, support vector machine, and neural network

Four machine learning models were evaluated to classify patient data: logistic regression, random forest, support vector machine (SVM), and neural networks. These models leveraged semantic features derived from the Word2Vec model, and additional features such as semantic density and voice similarity. Table 4 summarizes the OLS regression results, revealing that semantic density had a significant predictive impact on classification, whereas voice similarity was not as impactful.

Table 4. OLS regression results

Hyperparameters						
Entities	**Value**	**Entities**			**Value**	
Dependent Variable	Y	R-Squared (Uncentered)			0.875	
Model	OLS	Adjusted R-Squared (Uncentered)			0.875	
Method	Least Squares	F-Statistics			49.00	
No. Observations	16	Probability (F-Statistics)			4.77e-07	
Df Residuals	**14**	**Log-Likelihood**			**-0.52231**	
Df Model	2	AIC			5.045	
Covariance Type	Nonrobust	BIC			6.590	
Statistical Results of Independent Variable						
	coef	**Std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
x1	1.0000	0.169	5.916	0.000	0.637	1.363
x2	1.388e-16	0.158	8.78e-16	1.000	-0.339	0.339
Statistical Analysis						
Measured Entity	**Score**	**Measured Entity**			**Score**	
ANOVA	42.972	Durbin-Watson			2.000	
Probability (ANOVA)	0.000	Jarque-Beta (JB)			116.492	
Skew	3.615	Probability (JB)			5.06e-26	
Kurtosis	14.067	Cond. No.			3.02	

3.4.2. Calculation of semantic density and voice similarity

Semantic density and voice similarity calculations were incorporated into the predictive models. These metrics were based on the number of meaning components per content word and cosine similarity of voice-related words with each sentence. Table 5 outlines the comparison of these metrics between control and patient groups. Control data typically exhibited semantic densities above 0.81, while patient samples generally had lower densities.

Table 5. Results of semantic density and voice similarity

File	semantic_density	voice_similarity	expected
C1	0.848299	0.074384	0
C2	0.818362	0.089252	0
C3	0.825000	0.107641	0
C4	0.869258	0.096921	0
C5	0.939815	0.142358	0
C6	0.952381	0.091486	0
C7	0.837500	0.065277	0
C8	0.910570	0.102532	0
C9	0.812345	0.075678	0
C10	0.887654	0.094321	0
C11	0.825890	0.081234	0
C12	0.901234	0.107890	0
C13	0.935678	0.065432	0
C14	0.926789	0.085432	0
C15	0.819012	0.102345	0
P1	0.765842	0.116738	1
P2	0.730809	0.122966	1
P3	0.575488	0.101567	1
P4	0.852451	0.085209	1
P5	0.689677	0.109936	1
P6	0.509853	0.110761	1
P7	0.802779	0.130881	1
P8	0.595465	0.104845	1
P9	0.743526	0.111245	1
P10	0.781234	0.098765	1
P11	0.611234	0.113456	1
P12	0.826789	0.120345	1
P13	0.699012	0.108765	1
P14	0.527890	0.101234	1
P15	0.818901	0.129012	1

Voice similarity alone was insufficient to distinguish between the two groups, emphasizing the need for additional features.

3.4.3. Statistical validation of model performance using R-squared, F-statistics, and Mean Squared Error (MSE)

Statistical validation of model performance was achieved using metrics such as R-squared, F-statistics, and Mean Squared Error (MSE). The OLS regression model (Table 4) demonstrated strong predictive power, with an R-squared value of 0.875. The F-statistics value of 49.00 and a low associated probability underscored the model's significance. Further validation through ANOVA and the Jarque-Bera test confirmed residuals' normality and absence of autocorrelation. Despite promising results, the need for larger and more diverse datasets was highlighted to refine predictive accuracy.

The Durbin-Watson statistic, with a value of 2.000, suggests that there is no substantial evidence of autocorrelation in the residuals. Additionally, the values for skewness and kurtosis indicate deviations from normality and the presence of heavy tails in the residuals, respectively. The calculated condition number (Cond. No.) of 3.02 signals the potential presence of multi-collinearity among the independent variables. In summary, the results presented in Table 4 demonstrate the model's effectiveness in elucidating the dependent variable. Specifically, the variable 'x1' (Semantic Density) emerges as a statistically significant predictor, while the variable 'x2' (Voice Similarity) does not show any discernible impact. However, further research needed to address the issues associated with residual normality, multi-collinearity, and heavy-tailed residuals.

3.4.5. Hyperparameter tuning for predictive models

To conduct the training phase, each machine learning algorithm utilizes a variety of parameters. Examples include weights, the number of divisions (referred to as Estimators), the number of iterations (Epochs), and non-linearity functions

(Activation Functions). These parameters can be set in two ways: using hard-coded values or through hyperparameter optimization. In this study, we employed hard-coded values as they represent the default settings established through numerous trials. These were the most commonly used values for the parameters. Table 6 presents the hyperparameter values and their descriptions as used in this study.

Table 6. Hyperparameter values

Technique	Hyperparameter	Value	Description
Random Forest	n_estimator	50	The number of trees in the forest and the function to measure the quality of a split.
	bootstrap	False	Bootstrap Aggregation to select random samples from the dataset
	max_depth	110	The maximum depth of each decision tree in the forest
	max_features	1	The maximum number of features for each node
	min_samples_leaf	1	Minimum leaf nodes at each branch
	min_samples_split	5	The minimum number of samples required to split an internal node.
Support Vector Machine	C	0.1	The cost parameter of the SVM algorithm, the type of kernel function used to transform the input data, and the kernel coefficient.
	kernel	linear	The nature of the support line
Logistic Regression	penalty	l2	Regularization parameter that controls the trade-off between over-fitting and under-fitting.
	max_iter	300	Maximum number of repetitions allowed
	C	1	Inverse regularization strength.
	solver	'lbfgs'	Optimization method for weights. Usually works on the level of derivative applied in gradient descent
Neural Network	filters	16,32,64	The number of filters (also known as channels) used in the convolutional layers.
	kernel_size	3,5	The size of the convolutional kernel used to extract features from input data.
	activation function	Tanh	Converts the Function values using the Tanh function between -1 and 1
		Sigmoid	Converts the values between 0 and 1
		Softmax	Converts values based on the Entropy function
	Rate (drop)	0.25	Fraction of the input units to drop in the dropout layer.
	LR	0.01	The step size shrinkage is used in updating the weights of the functions at each layer.
	Optimizer	Adam	Algorithms used to change the attributes of neurons, for instance, weighs and learning rate to optimize the loss
	epoch	10	Number of iterations of the model
	Loss	Binary_crossentropy	Method for computing and returning loss value
	Metrics	accuracy	Model Evaluation metric

4. Outcomes

In summary, we aimed to predict the emergence of psychosis by analyzing speech patterns based on language features of low levels of semantic density and an increased tendency to talk about voices and sounds.

The first step was to identify low levels of semantic density involves pre-processing speech samples and extracting content words such as nouns, pronouns, verbs, and adverbs. Then, the sentence vector is computed by summing the vectors

of each content word and normalizing the result by its magnitude. The meaning component is derived from the vector unpacking process.

Finally, semantic density is calculated by dividing the number of meaning components by the number of content words, and the overall semantic density of the speech is obtained by taking the mean of each sentence's semantic density. In addition, the project also involved identifying an increased tendency to talk about voices and sounds by creating a list of voice-related words and computing their word vector by summing and normalizing them. The cosine similarity is then calculated between the voice-related words and the sentences, and the highest similarity is retained as an indicator of an increased tendency to talk about voices and sounds. The combination of these two indicators has shown a high level of accuracy in predicting future psychosis. Lastly, we demonstrated use of linear regression to train the models in this study and calculate accuracy.

5. Limitations and Recommendations

There were several limitations in this study. Firstly, there is a pressing need for a culturally appropriate psychometric screening tool to identify individuals with psychosis and or schizophrenia in indigenous population. A culturally appropriate screening questionnaire should contain questions that patients and families feel they can relate too.

In this research, tools such as BPRS and PANSS were abandoned, as patients could not relate to its questions. Secondly, the method of speech elicitation could be further refined to reflect a clinical discourse that has elements of eliciting symptoms of psychosis and schizophrenia or phenomenon of which indigenous patients could relate with ease.

Ideally conceptualising psychosis and or schizophrenia in terms of common themes such as Māori sickness versus mental illness may lend itself with ease in terms of patients and families engaging in a clinical discourse. The same could be applied to control samples of indigenous Māori whom are not known to have mental illness but would have shared understanding of the Māori worldview of mental illness. This would mean both patient and control sample will be exposed to speech and language tasks that is based on a universally understood cause of mental illness in indigenous Māori. As this can be viewed as an assessment of sociolinguistic determinants observed in psychosis and or schizophrenia, future research should focus on cognitive aspects of language, speech and abnormal forms in indigenous population using NLP methods. Conclusively this will assist further in assessing outcome measures when considering effectiveness of psychiatric treatment.

This study employed automated NLP techniques to detect linguistic impairments in individuals with schizophrenia or psychosis within a Māori population [22]. Unlike previous research, this study collected speech data from shared discourse in a clinical environment, without assigning specific speech elicitation tasks. It focused on observing speech-act behaviour and analysing semantic coherence. The results revealed that the patient group exhibited poorer semantic density and coherence compared to the control group, thereby confirming the hypothesis that language dysfunction serves as a marker for schizophrenia and psychotic disorders.

NLP tools used to mine for speech and language dysfunction in Māori language (Te Reo Māori) was not the focus of this study. This study examined speech and language disturbances in Māori patients whom spoke English. Linguistic and cultural nuances specific to Māori language and communication styles are vast as there are many different tribes in New Zealand. This means mining for speech and language dysfunction in patterns of indigenous communication requires multiple datasets of all Māori tribes across NZ. This study focused on one particular tribe in the Upper North Island of NZ. In such instances, it would make practical sense to train NLP tools in English language for improving early access to diagnosis and treatment.

Other limitations included the speech sample length; both patient and control samples averaged 10 minutes of speech. The study did not explore whether sample length influenced the reliability of total and different word counts in natural speech data collection. Furthermore, the data was not analysed to determine if the duration of speech samples affected semantic density and coherence.

Despite these limitations, each patient sample in this study contained more than 50 utterances, which is considered reliable for computing Latent Semantic Analysis (LSA) measures. The duration of these samples was measured in minutes. Although the findings support existing theories, the scarcity of studies in Indigenous populations may challenge the validity of the theory affirmed by this study, given its reliance on a very small sample and dataset. Mean Length of Utterance (MLU) is considered an indicator of language development [23]. Further research is needed to investigate how cultural factors, bilingualism, and multilingualism influence MLU and its manifestation in individuals considered at risk of developing psychotic illnesses.

This study used NLP to assess Māori patients (whom can be bilingual) who may or may not have diagnosis of psychosis or schizophrenia through speech and language analysis. The study analysed data samples in one language i.e. English. If other languages were used to collect speech data, then the proposed research papers such as Med-UniC are relevant [24].

It may be a consideration then that psychiatric language is considered a separate language to that of conversational English, which may vary with indigenous persons, as it will be influenced by a variety of other factors. There is a difference in terms of spoken English between native and non-native speakers. Indigenous Māori schooled in NZ are considered native English speakers.

In terms of frozen language and or multi-model assessment, it has no relevance to this study as no visual aids were used to elicit speech from patient samples. However, use of visual aids or texts that has cultural elements does shape form of communication and may have relevance in assessing for symptoms of schizophrenia [25]. Use of these other types or forms of language must be validated by cultural groupings requiring multi-site data gathering.

Further research using larger datasets may lend itself reliably with use of latent space geometry optimisation [26]. This model aims to compress data into lower dimensional space but retaining essential information. Due to heterogeneity of psychotic conditions inclusive of schizophrenia in a context of the cultural explanations, large datasets will be needed in order to use this NLP model for the purpose of reducing computational costs and improving effectiveness of data processing.

The challenge of small dataset in this study was tackled using synthetic data generations. Owing to reasons of privacy and confidentiality resulting in withdrawal of consent to participate in this study, there were a high number of dropouts resulting in the need to explore generation of synthetic data. In the chapter on methodology, challenges with data collection were discussed in detail and it was evident that privacy, confidentiality are possibly main barriers to a larger dataset. It was also noted that a longer period for data collection might have led to a larger dataset sample. Adequate length of sampling time may have resulted in some participants agreeing to have their data shared for the purpose of this research. Therefore, use of synthetic dataset can assist with increasing the chance of research and development.

Use of synthetic data have several challenges namely fidelity at the data level, fidelity at the population level, disclosure risks, generation of dataset close to original data distributions and preserve the attribute relations similar to the original data, most of the synthetic sampling technique address only one task for small dataset, regression and classification. There are limited methods for generating data for regression tasks.

While acknowledging the limitations of the current model based on a small dataset and the necessity for cross-validation with other indigenous Māori sub-groups in NZ, the study highlighted the potential of NLP to detect early indicators of schizophrenia in indigenous communities, based on features of speech dysfunction observed in schizophrenia. The study suggested that computational techniques for diagnosing schizophrenia based on linguistic anomalies could enhance early intervention and treatment. However, further research is required to test and refine the model's accuracy, incorporate cultural indicators and semantic attributes, and overcome challenges in data collection within indigenous patient populations.

The study underscored the importance of understanding Indigenous perspectives, collaborating with Indigenous scholars, and developing culturally appropriate psychometric tools for culturally responsive diagnostics. This research contributes to the field of NLP by focusing on linguistic irregularities in psychosis and schizophrenia, highlighting the critical need for precise and culturally sensitive healthcare methods.

This study's model successfully assimilated new Indigenous data and made accurate predictions despite the small sample size. This indicates that the model was able to generalize its training to new datasets. However, it was not cross validated with other Indigenous NZ subgroup samples, which is a limitation shared by many cross-cultural NLP studies on psychosis and schizophrenia speech abnormalities.

The practical implications of using this technology include the risks of over-diagnosis and misdiagnosis. However, the technology offers an individualized symptom profile model that can be assessed using NLP alongside treatment and other interventions. Further studies with larger samples are necessary to determine the validity and reliability of the results.

To some extent, the difficulties encountered with data collection for this project reflect the common challenges when comparing Māori health status to non-Māori health status in a healthcare setting that is not designed to meet the needs of Māori people. Therefore, extensive work is required to adjust these parameters in a healthcare environment that facilitates easy and early access to the assessment and early intervention.

6. Conclusions

In conclusion, our study helps address the gap in mental health assessments for Māori people. We attempted to minimize the language and cultural gaps in diagnostic procedures by investigating the ability of NLP to identify linguistic patterns suggestive of psychosis. The study emphasizes the value of inclusive language models that consider a variety of explanatory frameworks. Language classifiers, NLTK pre-processing, and the tripartite clinical interview technique all showed encouraging results. However, limitations stemming from a lack of data point to potential directions for future research, such as the use of deep learning methods to generate more comprehensive datasets. This research is a step toward improving the diagnostic accuracy and support for Māori people suffering from psychosis, as healthcare aims for fair and culturally responsive mental health screening.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

References

[1] Demarchi, C., Baune, B. T., & Clough, A. R. (2012). Detecting psychotic symptoms in Indigenous populations: a review of available assessment tools. Schizophrenia research, 139(1-3), 136-143. https://doi.org/10.1016/j.schres.2012.05.017

[2] Baxter, J. (2008). Māori mental health needs profile: A review of the evidence. Retrieved from: https://terauora.com/wp-content/uploads/2022/04/Māori-Mental-Health-Need-Profile-2008.pdf

[3] Carr, G., Cunningham, R., Petrović-van der Deen, F. S., Manuel, J., Gibb, S., Porter, R. J., ... & Lacey, C. (2023). Evolution of first episode psychosis diagnoses and health service use among young Māori and non-Māori—A New Zealand national cohort study. Early Intervention in Psychiatry, 17(3), 290-298. https://doi.org/10.1111/eip.13327

[4] Mellsop, G., Tapsell, R., & Menkes, D. B. (2019). Testing a hypothesis arising from the epidemiology of schizophrenia in New Zealand. General Psychiatry, 32(2). https://doi.org/10.1136/gpsych-2019-100048
[5] Kake, T. R., Garrett, N., & Te Aonui, M. (2016). Cognitive neuropsychological functioning in New Zealand Māori diagnosed with schizophrenia. Australian & New Zealand Journal of Psychiatry, 50(6), 566-576. https://doi.org/10.1177/0004867415607986
[6] Kake, T. R., Arnold, R., & Ellis, P. (2008). Estimating the prevalence of schizophrenia among New Zealand Māori: a capture–recapture approach. Australian & New Zealand Journal of Psychiatry, 42(11), 941-949. https://doi.org/10.1080/00048670802415376
[7] Allison, S., Bastiampillai, T., Castle, D., Mulder, R., & Beaglehole, B. (2019). The He Ara Oranga report: What's wrong with 'big psychiatry' in New Zealand?. Australian & New Zealand Journal of Psychiatry, 53(8), 724-726. https://doi.org/10.1177/0004867419848840
[8] Cunningham, R., Kvalsvig, A., Peterson, D., Kuehl, S., Gibb, S., McKenzie, S., ... & Every-Palmer, S. (2018). Stocktake report for the mental health and addiction inquiry. Retrieved from: https://mentalhealth.inquiry.govt.nz/assets/Summary-reports/Otago-stocktake.pdf
[9] Ratana, R., Sharifzadeh, H., Krishnan, J., & Pang, S. (2019). A comprehensive review of computational methods for automatic prediction of schizophrenia with insight into indigenous populations. Frontiers in psychiatry, 10, 659. https://doi.org/10.3389/fpsyt.2019.00659
[10] Ratana, R., Sharifzadeh, H., & Krishnan, J. (2023). Considerations for collecting data in Māori population for automatic detection of schizophrenia using natural language processing: a New Zealand experience. AI & SOCIETY, 1-12. https://doi.org/10.1007/s00146-023-01700-6
[11] Parola, A., Lin, J. M., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., ... & Fusaroli, R. (2023). Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence. Schizophrenia Research, 259, 59-70. https://doi.org/10.1016/j.schres.2022.07.002
[12] Srivastava, J., Sinha, V., Ketteler, D., & Jagtiani, A. (2022). High-Order Language Processing Difficulties in Patients With Schizophrenia: Cross-linguistic and Cross-cultural Results From the Hindi Version of a Newly Developed Language Test. Schizophrenia Bulletin Open, 3(1), sgac029. https://doi.org/10.1093/schizbullopen/sgac029
[13] Palaniyappan, L. (2021). More than a biomarker: could language be a biosocial marker of psychosis?. npj Schizophrenia, 7(1), 42. https://doi.org/10.1038/s41537-021-00172-1
[14] Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., ... & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. npj Schizophrenia, 1(1), 1-7. https://doi.org/10.1038/npjschz.2015.30
[15] Zandi, T., Havenaar, J. M., Limburg-Okken, A. G., Van Es, H., Sidali, S., Kadri, N., ... & Kahn, R. S. (2008). The need for culture sensitive diagnostic procedures: a study among psychotic patients in Morocco. Social Psychiatry and Psychiatric Epidemiology, 43, 244-250. https://doi.org/10.1007/s00127-007-0290-0
[16] Chaika, E., & Lambe, R. (1985). The locus of dysfunction in schizophrenic speech. Schizophrenia Bulletin, 11(1), 8-15. https://doi.org/10.1093/schbul/11.1.8
[17] Cohen, A. S., & Elvevåg, B. (2014). Automated computerized analysis of speech in psychiatric disorders. Current opinion in psychiatry, 27(3), 203-209. https://doi.org/10.1097/YCO.0000000000000056
[18] Singh, S., & Mahmood, A. (2021). The NLP cookbook: modern recipes for transformer based deep learning architectures. IEEE Access, 9, 68675-68702. https://doi.org/10.1109/ACCESS.2021.3077350
[19] Richardson, A. (2021). RE: Machine learning and medicine-three part series. Retrieved from: https://www.cmaj.ca/content/re-machine-learning-and-medicine-three-part-series
[20] Guillemin, M., Gillam, L., Barnard, E., Stewart, P., Walker, H., & Rosenthal, D. (2016). "We're checking them out": Indigenous and non-Indigenous research participants' accounts of deciding to be involved in research. International Journal for Equity in Health, 15, 1-10. https://doi.org/10.1186/s12939-016-0301-4
[21] Walter, M. (2005). Using the 'power of the data' within Indigenous research practice. Australian Aboriginal Studies, (2), 27-34. https://search.informit.org/doi/10.3316/informit.426079560974425
[22] Ngata, R. S. (2014). Understanding matakite : a Kaupapa Māori study on the impact of matakite/intuitive experiences on wellbeing. PhD Thesis, Massey University.
[23] Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). First language, 25(3), 365-376. https://doi.org/10.1177/0142723705059114
[24] Wan, Z., Liu, C., Zhang, M., Fu, J., Wang, B., Cheng, S., ... & Arcucci, R. (2024). Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. In Advances in Neural Information Processing Systems, 36.
[25] Chuang, C. Y., Lin, Y. T., Liu, C. C., Lee, L. E., Chang, H. Y., Liu, A. S., ... & Fu, L. C. (2023). Multimodal Assessment of Schizophrenia Symptom Severity from Linguistic, Acoustic and Visual Cues. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 3469-3479. https://doi.org/10.1109/TNSRE.2023.3307597
[26] Si, D., Cheng, S. C., Xing, R., Liu, C., & Wu, H. Y. (2019). Scaling up prediction of psychosis by natural language processing. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI),339-347. https://doi.org/10.1109/ICTAI.2019.00055

How to Cite: Ratana, R., Sharifzadeh, H., & Krishnan, J. G. (2024). Diagnosis of Schizophrenia and Psychosis in Māori People Using Speech Assay: A Natural Language Processing Approach. Medinformatics. https://doi.org/10.47852/bonviewMEDIN42022849