Proceedings of the 5th European International Conference on Industrial Engineering and Operations Management
Rome, Italy, July 26-28, 2022

Analyzing Lexical and Prosodic Features of Bilingual Job Interviews in Evaluating Interview Performance

Matthew Von Nathaniel Corcuera, Rudnick James Donaire, Rey Victor Mendillo, Dr. Larry Vea, and Joel De Goma
School of Information Technology
Mapúa University – Makati
Makati, Philippines
mvnecorcuera@mymail.mapua.edu.ph, rjadonaire@mymail.mapua.edu.ph,
rvdmendillo@mymail.mapua.edu.ph, LAVea@mapua.edu.ph, jcdegoma@mapua.edu.ph

Abstract

Different organizations have started to provide automation in evaluating applicants during their interviews. Due to globalization, bilingualism has become essential in assessing individuals. Past studies have focused on analyzing the features and developing models for evaluating applicants who spoke only one language – English – during their interviews. Hence, this paper aims to analyze the lexical and prosodic features of interviews spoken using both English and Tagalog in evaluating interview performances based on three criteria: Leadership Ability, Communication Skills, and Candidate Enthusiasm. After consulting with domain experts, 60 mock interview recordings were gathered. The researchers used Google Cloud Services API for transcription and translation, PRAAT to extract prosodic features, and LIWC to extract lexical features. The researchers gathered different algorithms from past studies and selected the most suitable ones based on the characteristics of the datasets. From the results of model training, the researchers concluded that the optimal algorithms were LDA for Leadership Ability with 94% accuracy and 95% F1-score, SVM for Communication Skills with 88.89% accuracy and 82% F1-score, and Random Forest for Candidate Enthusiasm with 87.5% accuracy and 89.2% F1-score. Furthermore, the researchers concluded that words categorized under Psychological Processes were the most prevalent for lexical features, while features under Frequencies and Formants were the most prevalent for prosodic features.

Keywords
Job Interviews, Bilingualism, Natural Language Processing, Multi-modal Behaviors, and Frequencies and Formants.

1. Introduction
An interview is a methodical approach in assessing an individual's capabilities (Huffcutt 2011). These capabilities vary based on what is being assessed (e.g., an individual's agreeableness, interpersonal skills, and confidence). It is a conventional method done by various organizations to hire applicants based on what it needs. Organizations rely more on job interviews to hire applicants as this has been common in assessing applicants' skills, personality, and capability to handle the job position (Daisuke et al. 2019).

Bilingualism has become essential in assessing individuals for employment. Ayada (2018) and Halai (2007) have stated that globalization has affected employment. As businesses and organizations start to shift to globalization, knowing two or more languages becomes one of the qualities employers are looking for in an individual. Another reason for this is the ability for bilingual individuals to express more information through the concept of code-switching (Lipski 1978).

Most of the conducted interviews in the mentioned studies are spoken in English; hence, this presents a gap because prosodic features of focus vary from one language to another (Lee 2015). Their models might not be effective when bilingual interviews are conducted. According to Ayada (2018) and Halai (2007), bilingualism is one of the qualities employers seek, as globalization affects the business industry. Thus, having a predictive model that can accommodate interview data of different languages would prove to be beneficial when it comes to automating how interview performances will be rated.

1.1 Objectives

The study aims to analyze lexical and prosodic features of bilingual job interviews in evaluating interview performance. Specifically, it aims to identify important lexical and prosodic features that affect the interviewee's performance in bilingual interviews. It also aims to explore the effects of removing momentary mistakes and personal information towards the performance of the model's prediction. Lastly, it also aims to determine which algorithm would provide the optimal results in evaluating interview performance with respect to the dataset and its features.

2. Literature Review

An interview is a methodical approach in assessing an individual's capabilities (Huffcutt 2011). It is a widespread practice done by various organizations to hire applicants based on several factors. These factors are measures on which interviewers rely in assessing an applicant. These measures range from different themes and focus subjects such as psychometric properties, interview formats, biases, and the like (Daisuke et. al 2019). These measures are then grouped into three: demographic characteristics, job-related content, and interviewee's performance (Huffcutt 2011).

Bilingualism refers to an individual's ability to speak two languages (Ayada 2018). These individuals can use two languages to create sensible and constructive sentences to form meaning. Globalization is a factor in how the world transitions in communicating with other countries because of several factors such as economics (Ayada 2018). With the increasing globalization, it is common to find people who use both English and their first language when communicating with each other (Halai 2007).

By definition, a lexicon refers to the component of a Natural Language Processing system, wherein semantic and grammatical information about individual words is found (Guthrie et. al 1996). Breaking this down further, a lexeme is a fundamental and meaningful unit in a lexicon (Merriam-Webster 2011). These lexicons, alongside prosodic cues, are among the features extracted by numerous studies such as the ones conducted by Das et al. (2017), Das et al. (2017), Gildea et al. (2018), and Joshi et al. (2018) to predict interview performance ratings. Natural Language Processing (NLP) is a field that aids computer machines in understanding the human language (Shruthi and Swamy 2020) which is done through several techniques such as tokenization, case folding, and lemmatization (Kurniawan et. al 2020).

Linguistic prosody is the rhythmic and intonational aspects of a language (Merriam-Webster 2011). These aspects are also known as suprasegmental, and they convey various communicative meanings that are not limited to emotions, speech act markings (such as assertions, requests, and questions), confidence, politeness, and even indexical functions (such as age, gender, and dialectal status of the speaker) (Prieto and Roseano 2018).

Several studies have attempted to create models to evaluate interview performances by predicting ratings for specific criteria. Many of these studies made use of extracted prosodic and lexical features. For instance, Gildea et al. (2018) presented a study that centralized on predicting the ratings for attributes found in interactions in job interviews. Das et al. (2017) also conducted a study that explored the application of using lexical and prosodic audio features in predicting social competency in job interviews. Their model provides summative feedback to candidates after an interview. Das et al. (2017) presented a model that automatically measures the communication skills of job candidates in behavioral interviews.

3. Methods

3.1 Data Preprocessing

3.1.1 Data Labelling

After gathering the mock interview recordings, these were then labeled based on the provided criteria by two annotators. Each interview recording was labeled, per criterion, using the Likert Scaling with the range of 1 to 5, with 5 being the highest score. Das et al. (2017) and Das et al. (2017) used a metric measurement known as Cohen's Kappa to get the inter-rater agreement between the raters.

3.1.2 Data Preparation

The audio recordings of the interviews were converted from .m4a format to .wav files using the FFmpeg library in Python in preparation for PRAAT and Google Cloud Speech-to-Text API to process the audio to lexical and prosodic features.

3.1.3 Lexical Features and Natural Language Processing

The Google Cloud Speech-to-Text API allowed the transforming of audio recordings of the interviews to transcriptions. Based on those transcripts, Natural Language Processing (NLP) was used to prepare the transcripts for Linguistic Inquiry and Word Counter (LIWC). These processes were doing case folding – where all texts were put to lowercases, tokenization – where it splits the transcript into sentences and words called *"tokens,"* and lastly, lemmatization – where it transforms words into its based forms.

LIWC consists of distinct categories of words that are based on their usage, such as positive and negative words, parts of speech (e.g., verbs, pronouns, adverbs), psychological words (e.g., social, cognitive, relativity), personal words, and other classes (e.g., fillers, assent). LIWC will provide the outputs of percentages of how many total words of each category are present in the transcript.

The researchers have used the LIWC 2007 dictionary to extract lexical features in transcripts. The categories present in LIWC 2007 have been utilized to extract lexical features in transcripts. These categories are Word Count, Summary Language Variables, Linguistic Dimensions, Psychological Processes, and Other Grammar.

3.1.4 Extracting Tagalog Lexical Features

In LIWC, there are only a limited number of supported languages, and Tagalog was not available. To capture the Tagalog words and apply LIWC, the process was to translate these words into English and apply LIWC. By doing this, the words initially spoken in Tagalog would be included in the list of features.
To measure the impact of translating Tagalog words to capture their lexemes in a bilingual interview, two sets of lexical features were extracted: one where Tagalog is considered as part of the lexical features, and second where Tagalog is considered as noisy data.

3.1.5 Prosodic Features

An open-source analysis tool, Praat, was utilized via Python's Parselmouth library to extract the prosodic features from the recordings of each interviewee. It is a computer software used for speech analysis in phonetics.
The features considered necessary for prosodic analysis include the pitch information depicted by F0, vocal intensities, characteristics of the first three formants of frequencies, and spectral energy. The features used in the study were divided into seven (7) distinct categories under Speech Analysis. These are spectral analysis, pitch analysis, formant analysis, intensity analysis, jitter, shimmer, voice breaks, cochleagram, and excitation pattern.

3.1.6 Data Augmentation and Balancing Class Distributions using SMOTE

The process of data augmentation is to create synthetic data points based on the original data points. The Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling approach that creates synthetic data points from the minority class samples (Blagus and Lusa 2013). In this study, SMOTE balanced the number of classes for each criterion to prevent imbalanced classification and add more samples to the datasets.

3.1.7 Standardization

The process of standardizing data refers to rescaling value distribution for making the mean and standard deviation of the observed values 0 and 1, respectively. This process was used before feeding the dataset into the selected classification algorithms, primarily when the inputs differed in scales. The Python function *standardize* can be obtained from the mlxtend. preprocessing package, which was used for standardizing the values of the dataset.

3.2 Feature Selection using Boruta Package

The Boruta package is a feature selection wrapper algorithm that iterates through a subset of features and identifies the relevant ones. It is done by adding random values in copied features (shadow features), training it in Random Forest Classifier to apply feature importance measures, and comparing if the feature's score is higher than the shadow feature. If the score of the original feature is higher than the shadow feature, it is an important feature. It removes features that scored lower in comparison to the shadow feature. It iterates this until all features are confirmed, rejected, or reach a specific iteration limit.

3.3 Building Models for Predicting Interview Performance Scores

The first step was to divide the datasets into 75% training sets and 25% testing sets using stratified random sampling. This was achieved through Python's sklearn library. The classification algorithms used to train the models were

decided upon from a list of algorithms used by past related studies after analyzing the features from the interview data to be gathered. This list of algorithms consists of Lasso Regression, Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Support Vector Regression (SVR), Decision Tree, Random Forest, and K-Nearest Neighbor (KNN).

The researchers used a set of criteria to analyze which of these classification algorithms will be selected to train the models. These criteria were developed based on the studies of Pise and Kulkarni (2016) and Ping and Xiong (2020), which includes the number of attributes or features in the dataset, the number of data instances or the sample size, the number of output classes, class entropy, and the type of data present in the dataset.

3.4 Testing Model Performance

After training, the models' performance was tested and gathered statistical metrics such as accuracy and F1-score. Accuracy is a well-known classification metric that can be used for either binary or multiclass classification problems. However, it would depend on the skewness of the interview data to determine whether accuracy will become a viable choice of evaluation. Apart from accuracy, the researchers also used F1-score, a harmonic mean of precision and recall. It is desirable to ensure that the predicted ratings for each interview criteria are the correct ratings (precision value) and for the models to capture as many correct predictions as possible (recall value).

3.5 Identifying Effects of Removing Momentary Mistakes and Personal Information

The lexical features representing momentary mistakes and personal information were removed to determine how removing momentary mistakes and personal information affects the model performance. These features are work, leisure, home, money, religion, death, swear words, netspeak, assent, non-fluencies, and fillers based on the category Psychological Processes in the LIWC.

3.6 Model Evaluation

As selecting the optimal models based on their performance results, the researchers have created a Python-based web application that will serve as a prototype for easy integration of models. On the website, a simple GUI functions will take inputs of an audio file in WAV format and the dictionary file taken from LIWC. On the other hand, as a substitute for long audio files that take much processing time, there is an option to upload the preprocessed features for demonstration purposes. The saved models are loaded and integrated into the website to perform the lexical and prosodic analysis of the recording. After the analysis, the candidate's interview performance scores for each criterion will be shown on the website.

4. Data Collection

The study required domain knowledge in Human Resource Management to identify the criteria as the basis for rating interview performances for an entry-level job position. The following were the criteria provided by the domain experts: Leadership Ability, Communication Skills, and Candidate Enthusiasm. A set of questions were also prepared for the mock interviews in which it was categorized into three: Verifying, Behavioral, and Skill-related questions.

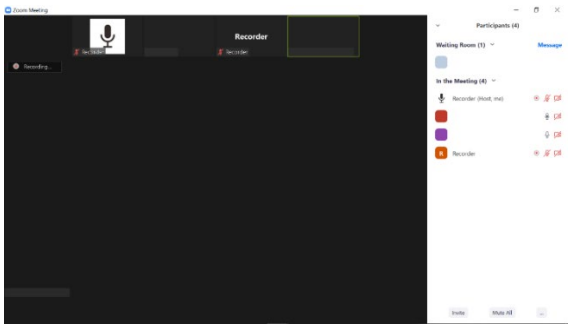

Figure 1. Mock Interview Setup

An online interview setup was followed to replicate a professional setting of a job interview. The interview setup was done in an online setting through Zoom, as shown in Figure 1. With the help of a professional HR interviewer, this

simulates the need for the interviews to be set in a professional setting. A total of 60 participants were gathered to be a part of the mock interview session. The participants were notified that both English and Tagalog were allowed when answering interview questions.

5. Results and Discussion

5.1 Data Preprocessing

Each interview recording was labeled by two annotators who have experience in conducting interviews with entry-level interviewees in the field of IT. The labels were based on the criteria that the domain experts have set, which are: Leadership Ability (LA), Communication Skills (CS), and Candidate Enthusiasm (CE). The resulting Kappa values to measure the agreeableness between the two annotators were in the range of .80 - .90, with an average of .90168, suggesting that the labels for each recording were highly agreeable.

5.2 Data Augmentation and Balancing Class Distribution using SMOTE

After applying SMOTE, results show that the class labels for each dataset were now balanced, and a slight increase in sample sizes was observed. Prior to data augmentation, all datasets had equal sample sizes of 60, with varying class labels distribution. Currently, LA has 125, CS has 105, and CE has 125.

5.3 Selected Features per Criterion

5.3.1 Lexical Features

After the data augmentation process, the researchers used the Boruta Package to select the essential and correlated features to be fed into the models. The lexical features shown above for each criterion are measures that can be grouped into four categories, as per the results of using LIWC explained in section 4.4 of the Methodology. Figures 1 to 3 show the percentage distribution of the features under these categories for each criterion.

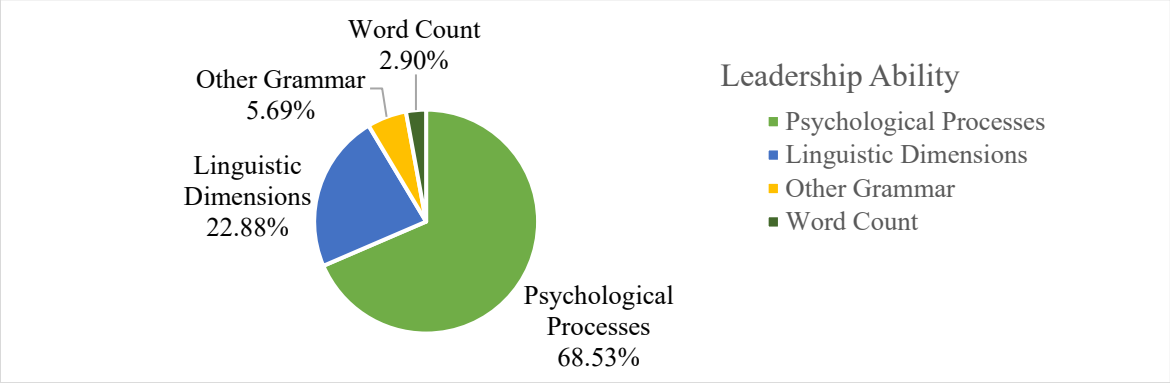

Figure 2. Lexical Features in Leadership Ability

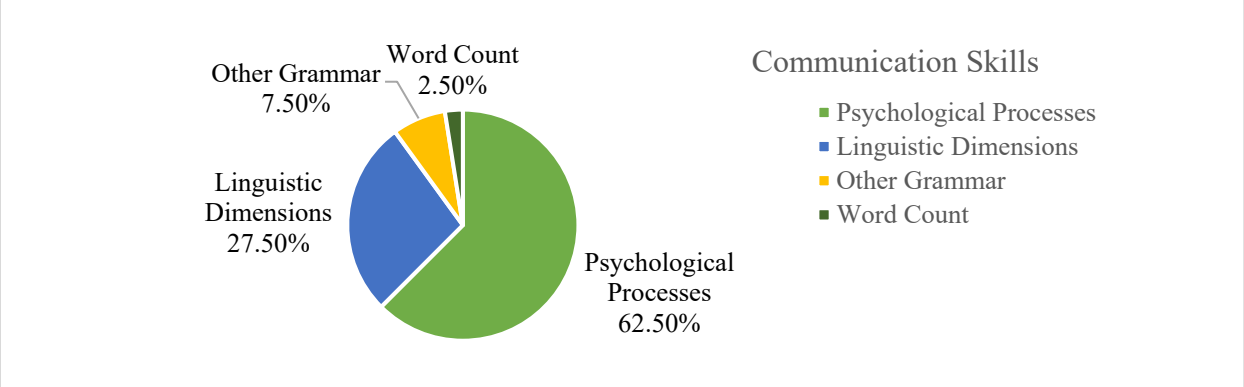

Figure 3. Lexical Features in Communication Skills

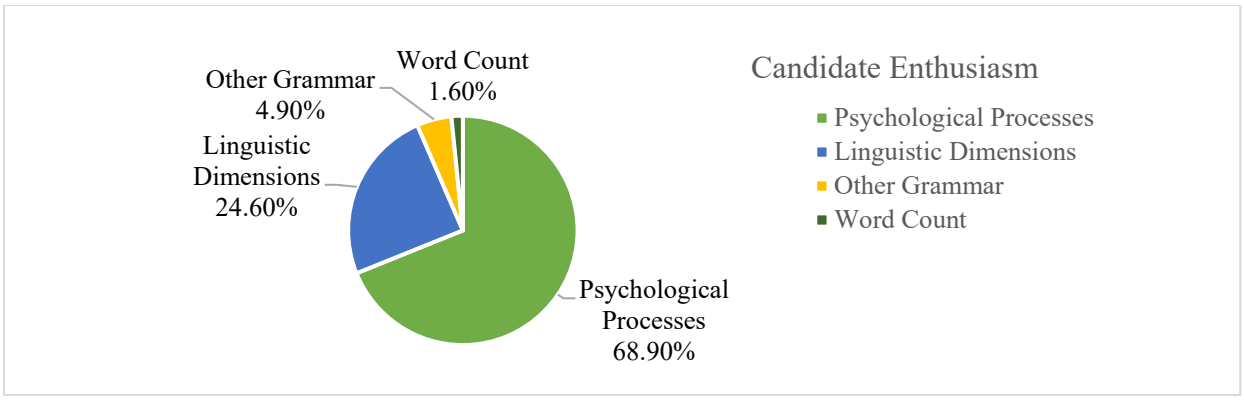

Figure 4. Lexical Features in Candidate Enthusiasm

Based on Figures 2 to 4, it can be observed that for all four criteria, most of the lexical features are those under the category of Psychological Processes. This is followed by the features under Linguistic Dimensions, Other Grammar, and Word Count, respectively.

5.3.2 Prosodic Features

Figures 5 to 7 show the percentage distribution of the features under these categories for each criterion. The prosodic features were grouped into their corresponding category of speech analysis based on their names to obtain these pie charts.

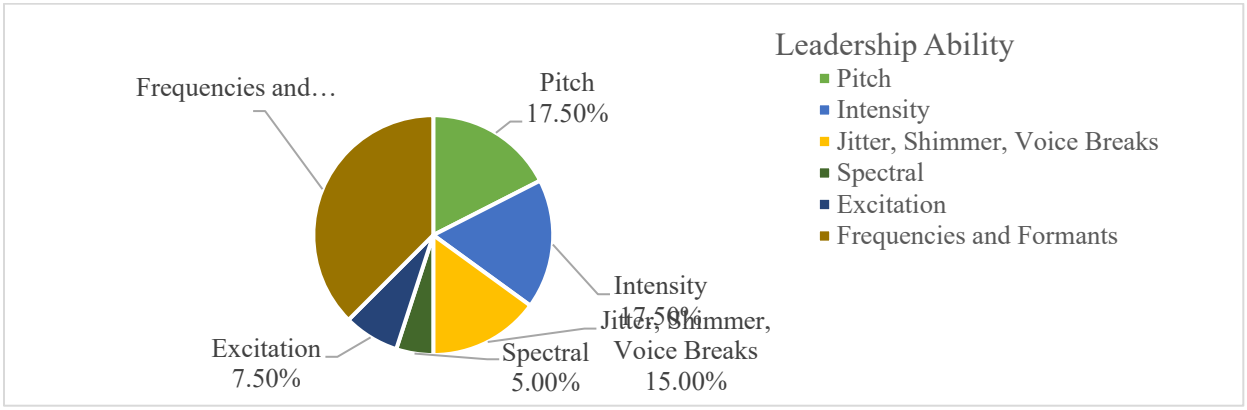

Figure 5. Prosodic Features in Leadership Ability

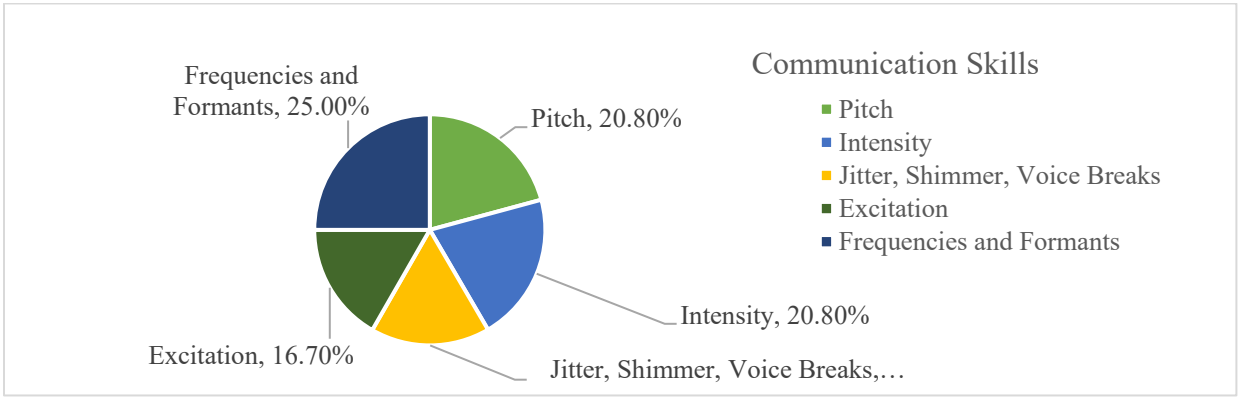

Figure 6. Prosodic Features in Communication Skills

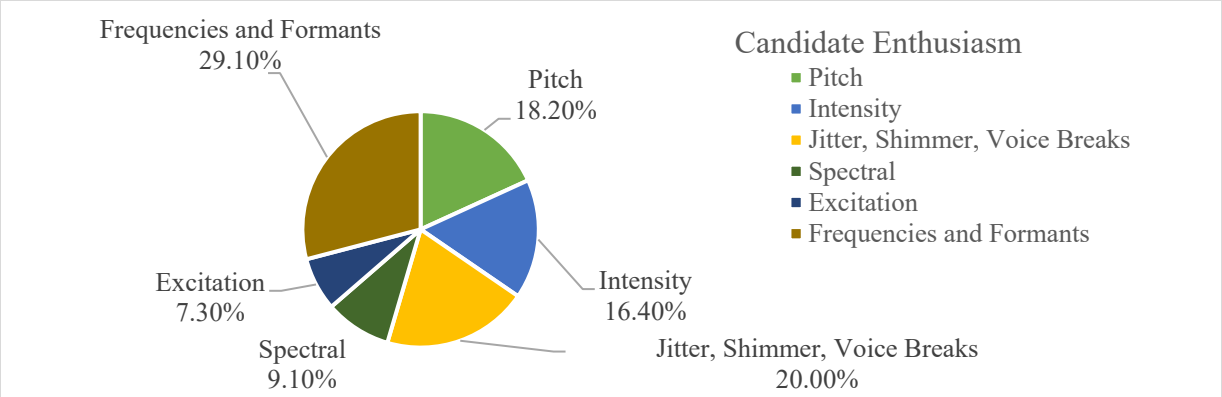

Figure 7. Prosodic Features in Candidate Enthusiasm

Based on Figures 5 to 7, most of the prosodic features for all criteria fall under Frequencies and Formants. Ranking the categories of these criteria in terms of distribution, Frequencies and Formants are first (29.1%), followed by jitter, shimmer, voice breaks (20.0%), then pitch analysis (18.2%). Next is intensity analysis (16.4%), then spectral analysis (9.1%), and finally excitation (7.3%).

5.4 Analysis of Features in Choosing Classification Algorithms

The dataset's characteristics were analyzed to determine the appropriate classification algorithms used in the study.

Table 1. Characteristics of Selected Features

Criteria	No. of Features	Sample Size	No. of Classes	Class Entropy	Type of Data
Leadership Ability	76	125	5	2.32	Numerical (Continuous)
Communication Skills	65	105	5	2.32	Numerical (Continuous
Candidate Enthusiasm	120	125	5	2.32	Numerical (Continuous)

Based on the characteristics of the feature shown in Table 1, the researchers have outlined several suitable classification algorithms corresponding to each criterion. The list of machine learning algorithms used for this analysis was collated from the algorithms used by past related studies such as Lasso Regression, Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Support Vector Regression (SVR), Decision Tree, Random Forest, and K-Nearest Neighbor (KNN).

Table 2. Suitable Algorithms

Criteria	Large No. of Features (> 50)	Large Sample Size (> 1030)	Multiclass	High-Class Entropy	Suitable Algorithms
Leadership Ability	✓	×	✓	✓	SVM, LDA, QDA, and Random Forest
Communication Skills	✓	×	✓	✓	SVM, LDA, QDA, and Random Forest
Candidate Enthusiasm	✓	×	✓	✓	SVM, LDA, QDA, and Random Forest

It can be observed in Table 2 that the characteristics of the features in each criterion are similar to one another, out of the nine different algorithms used by past the studies, SVM, LDA, QDA, and Random Forest.

5.5 Testing Model Performance Based on Accuracy and F1-Score

The suitable algorithms were used alongside the split training sets to train their respective classification models. The metrics of the algorithms per criterion, during and after model development are shown in Table 3.

Table 3. Metrics of each Algorithm per Criterion

Criteria	Algorithm	Train Accuracy	Train F1-Score	Test Accuracy	Test F1-Score
Leadership Ability	SVM	85.00%	85.00%	84.00%	85.00%
	Random Forest	94.62%	95.00%	87.50%	88.00%
	LDA	98.00%	98.00%	94.00%	95.00%
	QDA	83.00%	81.00%	84.00%	83.00%
Communication Skills	SVM	91.03%	91.60%	88.89%	82.00%
	Random Forest	83.33%	84.00%	81.48%	70.00%
	LDA	91.02%	92.00%	81.48%	67.00%
	QDA	98.72%	99.00%	77.78%	79.00%
Candidate Enthusiasm	SVM	89.24%	88.00%	84.38%	83.00%
	Random Forest	90.32%	90.00%	87.50%	89.20%
	LDA	83.87%	83.00%	75.00%	77.00%
	QDA	82.80%	81.60%	81.25%	82.40%

It can be observed that the optimal algorithm for predicting the ratings for Leadership Ability is LDA. On the other hand, the optimal algorithm for predicting the ratings for Communication Skills is SVM. Lastly, for Candidate Enthusiasm is Random Forest.

Table 4. Kappa Scores of Training Set

Criteria	Cohen's Kappa	Agreement Level
Leadership Ability	0.9597	Almost Perfect
Communication Skills	0.8878	Strong
Candidate Enthusiasm	0.8789	Strong

Table 5. Kappa Scores of Test Set

Criteria	Cohen's Kappa	Agreement Level
Leadership Ability	0.8441	Strong
Communication Skills	0.8606	Strong
Candidate Enthusiasm	0.8431	Strong

Measuring the model's inter-rater agreement against the actual data is analyzed. Both the training and testing datasets were used to analyze the agreement of the model. Table 4 shows that for training, the agreement levels for both CS and CE were "Strong," while LA had an agreement level of "Almost Perfect." On the other hand, Table 5 shows that across all criteria, the agreement level between the models and the labels was "Strong."

5.6 Results of Using Untranslated Dataset

The researchers created a set of data that detects Tagalog words as noise to discover its effects on the models' predictions for bilingual data. The data is then trained to a separate set of models.

Table 6. Model Performances using Untranslated Datasets

Criteria	Algorithm	Train Accuracy	Train F1-Score	Test Accuracy	Test F1-Score
Leadership Ability	SVM	67.00%	66.00%	41.00%	37.00%
	Random Forest	84.00%	82.00%	66.00%	58.00%
	LDA	92.00%	92.00%	55.00%	43.00%
	QDA	87.00%	86.00%	55.00%	50.00%
Communication Skills	SVM	95.06%	95.00%	66.67%	64.00%
	Random Forest	74.00%	71.00%	52.00%	45.00%
	LDA	99.00%	99.00%	67.00%	63.00%
	QDA	91.35%	91.00%	74.07%	73.00%
Candidate Enthusiasm	SVM	74.67%	74.00%	52.00%	51.00%
	Random Forest	75.00%	75.00%	48.00%	38.00%
	LDA	99.00%	99.00%	76.00%	75.00%
	QDA	99.00%	99.00%	80.00%	80.00%

As shown in Table 6, it can be observed that the optimal algorithm for predicting the ratings for Communication Skills, and Candidate Enthusiasm is QDA. On the other hand, the optimal algorithm for predicting the ratings for Leadership Ability is Random Forest.

After obtaining the optimal models for the untranslated dataset, their performances were then compared with those of the models trained with the translated dataset.

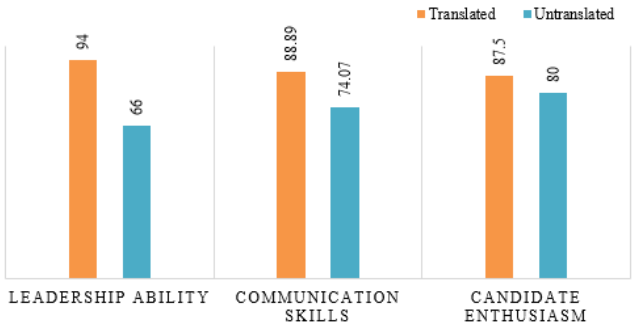

Figure 8. Optimal Model Comparison of Accuracies using Untranslated and Translated Datasets

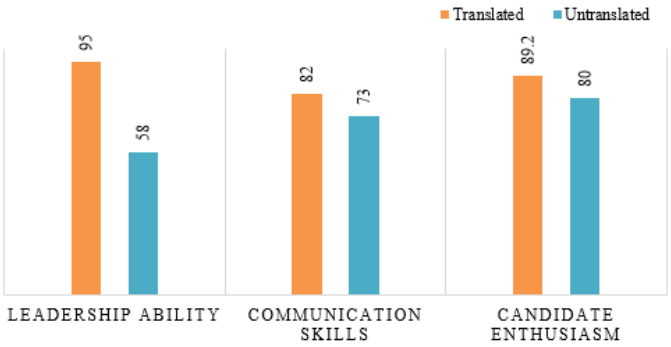

Figure 9. Optimal Model Comparison of F1 Scores using Untranslated and Translated Datasets

Based on Figures 5.8 and 5.9, the models trained using the translated dataset had better test accuracy and F1-score values when compared to the models trained using the untranslated dataset. With this, capturing the lexemes from Tagalog words through translation is vital for the models to have correct predictions on this study's bilingual interview data and better generalization.

5.7 Effects of Removing Personal Information and Momentary Mistakes

To address one of the objectives of this study, the researchers removed the features about personal information and momentary mistakes. This determines the effects of removing such features on the models' performances.

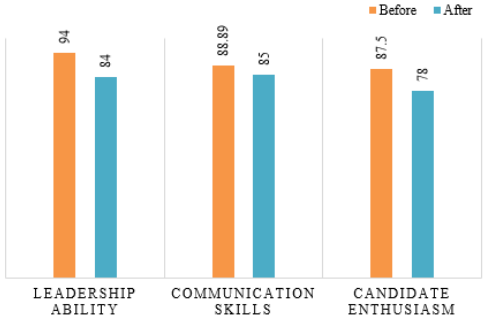

Figure 10. Effects on Accuracies in Removing Personal Information and Momentary Mistakes

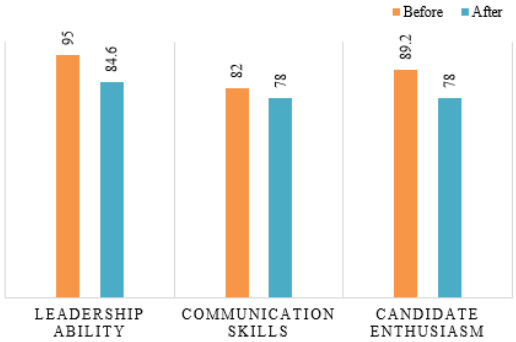

Figure 11. Effects of F1-Scores in Removing Personal Information and Momentary Mistakes

Based on the results shown in Figures 10 and 11, the model performances in terms of Accuracy and F1-Score for all criteria decreased after removing the features about personal information and momentary mistakes.

5.8 Model Evaluation using the Prototype

The researchers created a web application to integrate the optimal models to make a prototype. 6 new interview recordings from new interviewees were obtained and used in the Prototype. These were once again labeled with a set of scores for each interview criterion by an annotator. Using Cohen's Kappa, it was also measured to analyze the inter-rater agreement between the model and the annotator using actual data.

Table 7. Results of Kappa Validity test

Criteria	Cohen's Kappa	Agreement Level
Leadership Ability	0.67	Moderate
Communication Skills	0.5	Weak
Candidate Enthusiasm	0.7	Moderate

Table 7 shows that the agreement level between the Prototype's predictions and the domain expert's labels is Moderate

for the criteria Leadership Ability and Candidate Enthusiasm. On the other hand, the agreement level is Weak for the criterion Communication Skills.

6. Conclusion

For the first specific objective of the study, the researchers conclude that the essential features pertain to psychological processes and frequencies and formants. For the lexical features of all criteria, the most common lexical category was psychological processes, with a percentage distribution of 62% - 68%. As for the prosodic features of all criteria, formants and frequencies are prevalent, with 25% for CS, 37.5% for LA, and 29.1% for CE.

To measure the effect of translating Tagalog words into English to capture their lexemes for this study's bilingual data, a separate dataset was created that treated Tagalog words as noise. It was trained to a separate set of models. The results were compared to the metrics of those models trained with the inclusion of translated Tagalog words. The models that treated Tagalog words as noise showed poor accuracy and F1-score values and poor overall generalization due to the significant differences in their training and testing results. Therefore, capturing the lexemes from Tagalog words through translation is essential for the models to avoid overfitting, have similar predictions on the study's bilingual interview data, and have better generalization.

For the second specific objective, the researchers conclude that removing momentary mistakes and personal information negatively affected the performance of the models. This suggests that features about momentary mistakes and personal information are relevant to the candidate's overall performance during an interview.

For the third specific objective, the researchers conclude that the algorithms which provided the optimal results in predicting interview performance scores concerning this study's datasets and their features were LDA (for Leadership Ability), SVM (for Communication Skills), and Random Forest (for Candidate Enthusiasm). The Prototype's Kappa validity test results also show that using shorter interview recordings may result in predictions that exhibit weaker agreement levels. To further improve these models' predictions over unseen data, better convergence between training and testing performances, and stronger agreement levels in Kappa validity tests, adding more interview samples (varying in duration and context) to the datasets is preferable for future attempts.

References

Ayada, M., Bilingualism and Employability, 2018.

Blagus, R. and Lusa, L., SMOTE for high-dimensional class-imbalanced data, *BMC Bioinformatics*, vol. 14, no. 106, 2013.

Daisuke, O., Kiyoshi, T., and Norihiko, O., Interview Decision-Making by HR Practitioners: Statistical Policy-Capturing of Entry-Level Applicants, *International Conference on Research in Human Resource Management*, London, United Kingdom, March 7 – 9, 2019.

Das, R., Jayagopi, D.B., Nambiar, S.K., and Rasipuram, S., Automatic Generation of Actionable Feedback towards Improving Social Competency in Job Interviews, *Proceedings of 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, pp. 53 – 59, Glasgow, UK, 2017.

Das, R., Jayagopi, D.B., Pooja, R., and Rasipuram, S., Automatic Assessment of Communication Skill in Non-conventional Interview Settings: A Comparative Study, *Proceedings of 19th ACM International Conference on Multimodal Interaction*, pp. 221 – 229, New York, USA, 2017.

Gildea, D., Hoque, M.E., Naim, I., and Tanveer, I., Automated Analysis and Prediction of Job Interview Performance, *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 191 – 204, 2018.

Guthrie, L., Pustejovsky, J., Wilks, Y., and Slator, B.M., The role of lexicons in natural language processing, *Commun. ACM*, vol. 39, no. 1, pp. 63 – 72, January 1996.

Halai, N., Making Use of Bilingual Interview Data: Some Experiences from the Field, Available: https://www.researchgate.net/publication/265061773_What_are_parallel_and_comparable_corpora_and_how_can_we_use_them, April 25, 2021.

Huffcutt, A., An Empirical Review of the Employment Interview Construct Literature, *International Journal of Selection and Assessment*, vol. 19, no. 1, pp. 62 – 81, 2011.

Joshi, K., Kelhar, K., Kothari, V., and Sawla, H., Computational Inference of Candidates' Oratory Performance in Employment Interviews Based on Candidates' Vocal Analysis, *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, June 14 - 15, 2018.

Kulkarni, P. and Pise, N., Algorithm Selection for Classification Problems, *SAI Computing Conference*, pp. 203 – 211, London, UK, July 27 – 29, 2020.
Kurniawan, I., Lhaksmana, K., Richasdy, D., and Romadon, A.W., Analyzing TF-IDF and Word Embedding for Implementing Automation in Job Interview Grading, *2020 8th International Conference on Information and Communication Technology*, Yogyakarta, Indonesia, 2020.
Lee, Y. C., Prosodic Function Within and Across Languages, *Publicly Accessible Penn Dissertations*, United States, 2015.
Lipski, J., Code-switching and the problem of bilingual competence, *M. Paradis, editor, Aspects of Bilingualism*, pp. 250 – 264, 1978.
Merriam-Webster, 2011.
Ping, W. and Xiong, P., A Comparison of Classification Algorithms Based on the Number of Features, *Proceedings of the 39th Chinese Control Conference*, pp. 3179 – 3182, Shenyang, China, July 13 – 15, 2020.
Prieto, P. and Roseano, P., *Prosody: Stress, rhythm, and intonation*, Cambridge University Press, 2018.
Shruthi, J. and Swamy, S.K., A prior case study of natural language processing on different domains, *International Journal of Electrical and Computer Engineering*, vol. 10, no. 5, pp. 4928-4936, October 2020.

Biographies

Rey Victor D. Mendillo holds a Bachelor of Science degree in Computer Science with specialization in Artificial Intelligence from Mapúa University. In mid-2021, he worked at Chimes Consulting as Associate Back-End Developer, where he deployed custom user roles and e-mail automation with notifications in its Accounting and HR systems and assisted in resolving test cases related to user experience. His research interests concern machine learning, computer vision, natural language processing, numerical analysis, optimization algorithms, and automation.

Matthew Von Nathaniel E. Corcuera holds a Bachelor of Science degree in Computer Science from Mapúa University and specializes in Artificial Intelligence. Apart from his knowledge in A.I. and machine learning fields, he has also had hands-on experience with Robotic Process Automation (RPA) during his internship with Robinsons Bank Corporation, where he developed a software robot that automates sending of Creditable Withholding Tax emails to respective clients. He currently works as an Associate Technical Consultant in Infor PSSC Inc., and his areas of interest include artificial intelligence, machine learning, RPA, functional programming, and web design.

Rudnick James A. Donaire holds a Bachelor of Science degree in Computer Science specializing in Artificial Intelligence from Mapúa University. In his college years, he has gained knowledge about Machine Learning, Artificial Intelligence, and Data Science. In late-2021, he worked under a Professor in Mapúa University doing research about Information Systems and preparing lecture materials. He is mostly interested in the field of Artificial Intelligence and Data Science and is currently making personal projects such as analyzing personal data to develop highly needed skills.

Dr. Larry A. Vea is a former professor from Mapúa University. He has taken PhD in Computer Science in Ateneo de Manila University, Master of Science in Computer Science in De La Salle University – Manila, and Bachelor of Science in Electrical Engineering in Mariano Marcos State University. One of his latest publications is, “Real-Time Human Sitting Position Recognition using Wireless Sensors.”

Joel C. De Goma is a professor from Mapúa University. He has taken Master of Science in Computer Science in Mapúa University, Master of Engineering in Electronics Engineering in Mapúa University, and Bachelor of Science in Electronics Engineering in Mapúa University. One of his latest publications is, “Detecting Red-Light Runners (RLR) and Speeding Violation through Video Captures.”