Autonomous Language Processing and Text Mining by Data Analytics for Business Solutions

Voon Hee Wong, Wei Lun Tan(✉), Jia Li Kor, and Xiao Ven Wan

Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
tanwl@utar.edu.my

Abstract. Speech analytics solution is a technology that enables a company to discover customer's patterns and insights by analyzing relevant data, such as recorded audio files or phone conversations. The accuracy of speech recognition or speech-to-text transcription has been a challenge all along. This paper aims to present a text classification model for the call transcriptions based on the context, and to improve the accuracy of Google Speech API in Malay language. In this study, the accuracy of speech-to-text transcription is measured by word recognition rate and an accuracy scale. Time-cut-point and audio speed are the factors investigated to determine whether these factors affect the accuracy of text transcription. The results obtained from different time-cut-point and audio speed setting have been studied to identify the best combination. Furthermore, the pre-processed text data is utilized to train the text classification model using Support Vector Machine and Naive Bayes algorithms. In this paper, two approaches have been studied to improve Google Speech API. The first approach is to apply speech adaptation, which is the function made by Google. However, it showed that the accuracy dropped when 250 words were added into the speech adaptation, or when the audio speed was lowered. This is because the words error rate for both methods have increased. In the second approach, removing speech adaptation and lowering audio speed simultaneously caused a decrease in words error rate, hence the accuracy increased. In a nutshell, Support Vector Machine has better accuracy score of text classification as compared with Naive Bayes algorithms. As a result, short time-cut-point with normal speed of audio file showed a positive impact to improve Google speech-to-text API, along with Support Vector Machine being more suitable for classification model.

Keywords: Speech analytics solution · Word recognition rate · Support vector machine · Naïve Bayes algorithms

1 Introduction

Speech analytics solution was built by the combination of speech recognition and natural language processing (NLP). Speech-to-text is a software that can transcribe the voice into text. This software had become popular worldwide, especially for those who need to generate content with long descriptions automatically, and eventually increased working

N. Wahi et al. (Eds.): ICMSS 2022, ACSR 98, pp. 85–93, 2023.
https://doi.org/10.2991/978-94-6463-014-5_9

efficiency and productivity. NLP helped a computer to understand languages spoken by humans. It was explained as an automated way of analysing the written text by following some theories and technologies. The combination of NLP and speech recognition increased accuracy by training a model using machine learning. The model functioned to figure out and understand a topic rather than match a group of words. We studied the transcription's inaccuracy in detecting the Malay and mixed languages are spoken by Malaysians. Speech analytics solutions are always popular in call centers. Call centers play an essential role in a company as it provides a critical connection in the value chain that link their client with the company. To improve customer satisfaction in the customer service field, speech-to-text (STT) and text mining help make the work done efficiently. The combination of STT and text mining can help the company in several ways to improve business efficiency, e.g., classify call recordings based on the calling purpose. Furthermore, speaker diarization is a process that divides a multi-person audio stream into uniform parts associated with each person [1].

Speech analytics did analyse and extract the information from the unstructured audio data. Originally, speech analytics was known as audio analytics, and it transformed its name into speech analytics when it applied to the languages human's speech [2]. The general terms of speech recognition were a human-computer interface, speech processing, pattern recognition, modelling technique and signal processing. Speech recognition had four working stages: analysis, feature extraction, modelling, and testing or matching [3]. Google Speech Recognition was introduced in the year 2008. Google has stored huge quantities of data and some machine learning algorithms in its server. With large data and ready algorithms, Google developed its first large scale Automatic Speech Recognition (ASR) system. In 2017, Google Cloud Speech to Text API (Google STT API) was introduced to the public. Google STT API could perform ASR, but it also provided other extra features such as real-time streaming and automatic punctuation and auto-detection of languages [4].

Speech recognition performance was evaluated by calculating its accuracy using Word Recognition Rate (WRR). WRR had some limitations in that it neglected the words' importance and evaluated all the correctness in a document with the same score. This was not applicable in the real world as the words were important because some were the keywords in a transcription. Besides, WRR also ignored the speaker labels. For example, the first sentence was spoken by person A, and the second sentence was spoken by person B; in the way of calculating WER, we would not match the sentence with the person but only focused on counting the words recognized correctly [5]. Ordenes [6], defined text mining as analysing text data to explore their structure, pattern, and "hidden" meaning within the text. By mining textual data for call transcriptions, businesses could enhance their decision making for better resources allocation and better marketing strategy for their products. Text mining has been broadly utilized in different e-mail filtering, report, or document management [7].

The objective of this study is to determine the factors that affect the accuracy of the Google Speech-to-text Application Programming Interface on the text transcription in Malay and mixed language, which is mixture of Malay and English, with the use of speech analytics solution and to investigate the best combination of the various time duration and audio's speed that gives the highest accuracy of the transcription. Besides, this study also aims to increase the accuracy for Google Speech-to-text Application Programming Interface for Malay and mixed language and build a text classification model for the call transcriptions based on their categories.

2 Data

There are only 30 call recordings in Waveform Audio File format being analysed and used to figure out the best combination of time duration and speed that gives the highest accuracy. All the call transcriptions are generated by Google STT API and printed in an excel file. 50 call recordings have applied speaker diarization to analyse their accuracy improvement by speech adaptation and audio speed adjustment. While a total of 306 call transcriptions were used to train the topic classification model.

3 Methodology

3.1 Factors Affecting Accuracy and Combination of Time Duration(s) and Speed(s)

This study was aimed to increase the accuracy of the speech analytics solution that could convert spoken words into written texts. Quantitative and secondary data were used, and they are call recordings related to insurance policy cancellation, which were being analyzed and used as the training data. The Google Speech to Text (Google STT) API has been approached to transcript the call recordings into text. The Google STT API has been used because of the high accuracy and various features on transcript the speech into text. As Google STT API recognizer had a limitation that the audio's duration could not be more than one minute, thus Pydub module in Python was used to trim the audio files. After the conversion, we validated the text transcription's accuracy by listening and compared the text transcription with the call recordings manually.

Speech recognition performance was evaluated by Word Recognition Rate (WRR). The WRR was used to measure the accuracy of the text transcription [8]. The mathematical formation of WRR is modelled as Eq. (1). The higher value of the WRR indicated the higher performance and accuracy of the text transcription.

$$WRR = \frac{N - S - D - I}{N} = \frac{H - I}{N} \tag{1}$$

where S = Number of substitutions, D = Number of deletions, I = Number of insertions, N = Number of words and H indicates the number of words that are recognized correctly. In addition, the accuracy scale table is designed in assessing the accuracy of the text transcription. Table 1 showed the accuracy scale on the criteria of the "Text Understanding", "Wrong Spelling of Word" and "Completeness of Transcription" from the transcription.

Table 1. Accuracy Scale Table

Scale parameter	Accuracy description	Text understanding	Wrong spelling of word	Completeness of transcription
1	Very low	<20%	>20 words	<20%
2	Low	20%–<40%	15–20 words	20%–<40%
3	Medium	40%–<60%	10–15 words	40%–<60%
4	High	60%–<80%	5–10 words	60%–<80%
5	Very high	>80%	1–5 words	>80%

3.2 Speaker Diarization with Speech Adaptation

Another method to increase the accuracy is applying speech adaptation when performing speech recognition and reducing the audio speed. Speech Adaptation is a feature provided by Google STT API to improve their speech-to-text API by adapting 250 words, including the Malay common words and common insurance terms. Besides, reducing the audio to 0.75× is one of the methods to increase the accuracy of speech recognition. There are 51 call transcriptions under "Product Enquiry" category that would be tested in this study. On the other hand, the accuracy of the speech recognition would be determined by the WRR.

3.3 Topic Classification Model Development

There are two classification algorithms used in this study which are Naive Bayes which its definition function is explained in (2), and the algorithm of Support Vector Machine is explained in (3). Naive Bayes methods are based on Bayes' Theorem and the "Naive" assumption of conditional independence between every pair of features given the value of the class variable.

$$\widehat{y} = \arg \underset{y}{max}\ P(y)\prod\nolimits_{i=1}^{n} P(x_i|y) \tag{2}$$

where n is the number of variables, y is the class variable, x_i until x_n is the dependent feature vector [9].

Linear Support Vector Classification (Linear SVC) and Support Vector Classification (SVC) are two major forms of SVC that can do multiclass classification. SVC can be implemented more quickly in the linear kernel.

$$\sum\nolimits_{i=1}^{n} y_i a_i K(x_i,\ x) + b \tag{3}$$

where x_i is the training vector, y is the vector, $y \in \{1,\ -1\}^n,\ a_i$ is the dual coefficient which upper bounded by C, $K(x_i,\ x) = \phi(x_i)^T \phi\big(x_j\big)$ is the kernel [10].

Besides, the model evaluation scores obtained from Scikit-learn library also represent the models' accuracy score in text classification. The accuracy score can be calculated as (4).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

where TP is total positive which denotes the number of positive class documents categorised by the system, TN is total negative which indicates the number of negative class documents correctly identified by the system, FN is false negative which represents the number of positive class documents incorrectly classed by the system, and FP is false positive which signifies the number of negative class documents incorrectly categorised by the system [11].

4 Discussion

4.1 Factors Affecting Accuracy and Combination of Time Duration(s) and Speed(s)

The transcription accuracy is analysed on difference time duration or called as difference "Time Cut Point". Four different time durations are applied on the call recording audio which are 59 s, 20 s, 10 s, and 5 s. The unit of time being used in Python programming was in milliseconds, meaning 1 s would be inserted as 1000 ms. Besides, the transcription accuracy also analysed on the speed of audio. Five different speeds were being tested, including 0.5× (0.5 times slower than normal speed), 0.75×, 1× (normal speed), 1.25× and 1.5×. Both factors (Time duration and speed) are analysed together in a combination form, for example, 5 s with 0.5×, 5 s with 0.75× and so on. Each call recording audio file had 20 combinations. There were a total of 30 call recording audio files with different combinations that are investigated.

Figure 1 showed the total number of audio files for each combination when they performed the accuracy evaluation on the reference to the accuracy scale table and Word Recognition Rate (WRR). A total of 16 out of 30 audio files from the accuracy scale had achieved the highest accuracy with the combination of a time duration of 5 s and normal speed audio. Similarly, a total of 14 out of 30 audio files from the WRR shown the best result with the combination of a time duration of 5 s and normal speed audio. It can be seen that, more than 50% of the samples have achieved higher accuracy in the combination of 1× and 5 s in both accuracy scale table and WRR. We can conclude that the audio file performed better when it was being cut for every 5 s at normal speed among all the combinations.

4.2 Speaker Diarization with Speech Adaptation and Speed

Table 2 showed the statistics summary for 50 call transcriptions on original, speech adaptation and speech adaptation associated with 0.75× speed. The average of WRR for 50 original call transcriptions is decreased from 41.39% to 36.77% while 250 words applied into speech adaptation function from Google STT API. It is indicated that the accuracy has been dropped 3.09% for 250 words adaptation. The average of WRR is decreased from 36.77% to 35.56% when additional step to make audio to 0.75× speed.

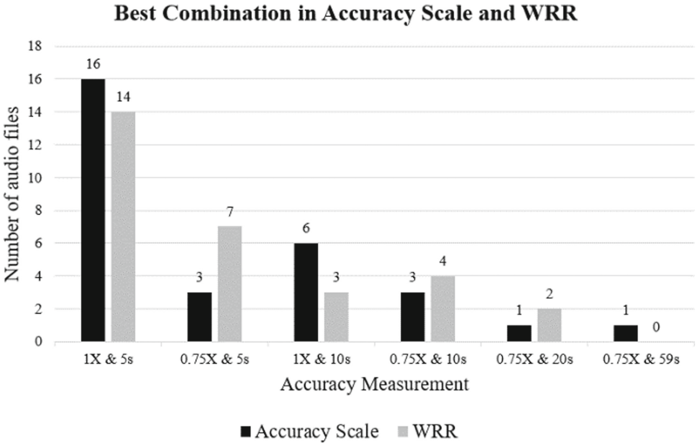

Fig. 1. The number of Audio Files that Obtained Best Accuracy in Different Combinations in Each Accuracy Measurement.

Table 2. The descriptive statistical data analysis for original transcription, 250 words adaptation transcription and 250 words adaptation with 0.75× speed.

Methods	Word Recognition Rate (WRR)			
	Average	Quartile 1	Median	Quartile 3
Original Transcription	0.4139	0.3601	0.4213	0.4653
250 Words Adaptation	0.3677	0.3194	0.3822	0.4308
250 Words Adaptation + 0.75× speed	0.3556	0.3063	0.3713	0.4241

This show that the speech adaptation cannot improve the average accuracy of the speech to text on Bahasa Malaysia transcription in term of WRR. Figure 2 showed the boxplot for WRR on three difference methods. From the distribution of WRR on original and both pre-processed audios, the pre-processing audio unable to provide more accuracy transcription.

Table 3 showed the statistics summary for 50 csall transcriptions on original, 0.75× speech transcription and speech transcription with 0.75× speed and clear audio transcription. The average of WRR for 50 original call transcriptions is increased from 41.39% to 45.33% while slower the audio speed for call recordings before performing speech recognition. It is indicated that the accuracy has been improved 3.09% by slower down the audio speed transcriptions. However, the average of WRR is decreased from 45.33% to 44.69% when additional step to make audio to be clearer by audacity software. Hence, the average accuracy of the speech to text on Malay and mixed language in term of WRR has been increased while lowering the audio speed to 0.75×, but the accuracy declines 0.64% while the audio file is being pre-processed to be clearer. Figure 3 showed the boxplot for WRR on three difference methods. The range of the WRR on both pre-processed audios has slightly increase compared to original audios. It can be concluded that, the

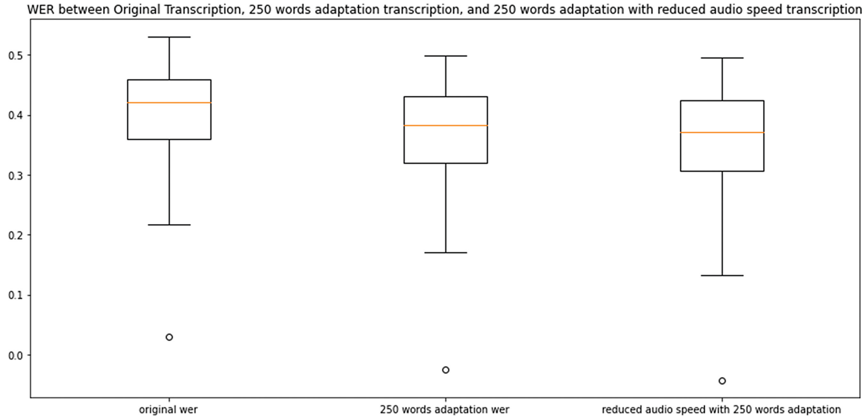

Fig. 2. WRR between original transcription (left), 250 words adaptation transcription (middle) and 250 words adaptation with reduced audio speed transcription (right).

Table 3. The descriptive statistical data analysis for original transcription, 0.75× audio speed transcription and clearer audio with 0.75× speed transcription.

Methods	Word recognition rate (WRR)			
	Average	Quartile 1	Median	Quartile 3
Original Transcription	0.4139	0. 3601	0.4213	0.4653
0.75× Speed Transcription	0.4533	0.3949	0.4651	0.5226
0.75× Speed With Clearer Audio Transcription	0.4469	0.3886	0.4628	0.5180

audio after pre-processing can obtain more accuracy transcription compared to original audio.

4.3 Assessment on the Model Classification

The machine learning algorithm of Naive Bayes and SVM are used to train the text classification model based on the calling purpose. It is critical in assessing the model's performance to determine the best algorithm for text classification. The model predict ability has been assessed through the accuracy score. First, the textual data is split into 70% as a training data set and 30% as a testing data set. The training data set are used to train the classification model to make the predictions. After training, the model used the remaining 30% of the dataset to compute the predictions of accuracy score. Table 4 showed the accuracy scores for Naive Bayes and support SVM.

From Table 4, the accuracy scores for Naive Bayes algorithm is 52.90% while for SVM algorithm is 67.10%. The difference between these two classifiers is 14.19%. It carries the meaning that SVM has a better prediction result for text classification in this study compared to Naive Bayes.

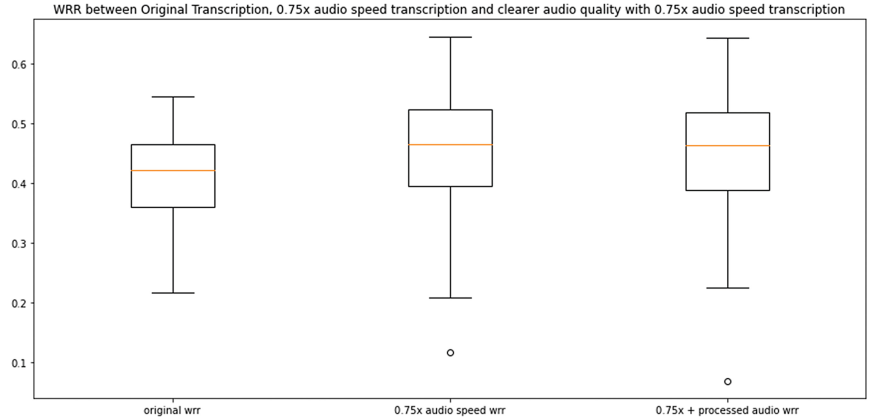

Fig. 3. WRR between original transcription (left), reduced audio speed transcription (middle), and reduced audio speed with clearer audio quality transcription (right).

Table 4. The accuracy score for the testing set from Naive Bayes and Support Vector Machine algorithms.

Algorithm	Accuracy Score
Naive Bayes	0.5290
Support Vector Machine	0.6710

5 Conclusion

The factors that affect the accuracy of the Google STT API on the text transcription in Malay and mixed language have been identify and the text classification model for the call transcriptions have been determined. The combination of the time duration and audio speeds have been applied to obtain the highest accuracy on text transcription. The combination of 1× audio speed and 5 s is the best combination for both accuracy scale and WRR. The results show that more than 50% of the samples achieved the highest accuracy with this combination. In the other hand, the way to increase the accuracy of speech recognition for Malay and mixed language is to lower the audio speed to 0.75× . In the text classification model, support vector machine (SVM) classifier performs the best compared to the Naive Bayes classifier.

References

1. Wang, Q., Downey, C., Wan, L., Mansfield, P.A., Moreno, I.L.: Speaker diarization with LSTM. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5239–5243. IEEE (2018)
2. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**(2), 137–144 (2015)

3. Gaikwad, S., Gawali, B., Yannawar, P.: A review on speech recognition technique. Int. J. Comput. Appl. **10**(3), 16–24 (2010)
4. Iancu, B.: Evaluating google speech-to-text API's performance for romanian e-learning resources. Informatica Economica **23**(12019), 17–25 (2019)
5. Balestrieri, E., Catelani, M., Ciani, L., Rapuano, S., Zanobini, A.: Word error rate measurement uncertainty estimation in digitizing waveform recorders. Measurement **46**(1), 572–581 (2013)
6. Ordenes, F., Theodoulidis, B., Burton, J., Gruber, T., Zaki, M.: Analyzing customer experience feedback using text mining. J. Serv. Res. **17**(3), 278–295 (2014)
7. Ur-Rahman, N, Harding, J.: Textual data mining for industrial knowledge management and text classification: A business oriented approach Expert Systems with Applications, vol. 39(5), pp. 4729–4739(2012)
8. Ogawa, A., Hori, T.: Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. Speech Commun. **89**, 70–83 (2017)
9. Kolluri, J., Razia, S.: Text classification using Naive Bayes classifier. Mater. Today: Proc. (2020)
10. Chauhan, A., Agarwal, A., Sulthana, R.: Genetic algorithm and ensemble learning aided text classification using support vector machines. Int. J. Adv. Comput. Sci. Appl. **12**(8), 260–267 (2021)
11. Lutfi, A., Permanasari, A., Fauziati, S.: Sentiment analysis in the sales review of indonesian marketplace by utilizing support vector machine. J. Inf. Syst. Eng. Bus. Intell. **4**(1), 57 (2018)