iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58631

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Sounding Sentiments: A Cognitive AI Approach to Speaker-Linked Emotion Analysis

Dr. Seema[1], Gungun Tyagi[2], Kumar Divij[3], Amitesh Maity[4], Annu Redhu[5], Srishti[6]
Dept. of CSE, Chandigarh University

Abstract*: Over the past few decades, there has been significant progress in sentiment analysis, primarily focusing on analyzing text. However, the field of sentiment analysis linked to audio remains relatively undeveloped in the scientific community. This study aims to address this gap by introducing sentiment analysis applied to voice transcripts, specifically focusing on distinguishing emotions of individual speakers in conversations. The proposed research article seeks to develop a sentiment analysis system capable of rapidly interacting with multiple users and analyzing the sentiment of each user's audio input. Key components of this approach include speech recognition, Mel-frequency cepstral coefficients (MFCC), dynamic time warping (DTW), sentiment analysis, and speaker recognition.

I. INTRODUCTION

Sentiment analysis is used to analyze people's feelings or attitudes based on a conversation, topic, or general conversation. Sentiment analysis is used for various purposes such as in various applications and websites. We use our knowledge to create an assistant that understands and learns the human way of thinking based on holding each other. A machine that understands people's emotions/moods through these conversations and what keyword was used in the conversations. The combined sentiment analysis of the speaker and the speech is done with data and conversations extracted from previous conversations and various processes.

Understanding people's thoughts and feelings has many applications. For example, technology that can understand and respond to a person's personal emotions will be important. Imagine a device that senses a person's mood and adjusts its settings based on the user's preferences and needs. Such innovations can improve user experience and satisfaction. In addition, research institutions are actively working to improve the quality and translation of audio content into text. This includes a variety of materials such as news reports, political speeches, social media, and music. By enabling this technology, they aim to make audio content more accessible and useful in many situations.

Our research body has also worked on voice evaluation research [1,2,3] to study the conversation between the user and the assistance model and distinguish between each speaker and their emotions. Because there are several speakers in a conversation, it is difficult to analyze the text data of the recorded voice, so this paper proposes a model that can easily recognize the presence of different speakers and identify them as separate, and perform voice analysis; individuals speak and responds according to his feelings.

We present an approach and viewpoint on investigating the hurdles and techniques involved in audio perception analysis of sound recordings through speech recognition. Our methodology utilizes a speech recognition model to interpret audio recordings, coupled with a suggested method for user discrimination based on a predetermined hypothesis to authenticate distinct speakers. Subsequently, each segment of speech data is scrutinized, enabling the system to accurately discern genuine emotions and topics of conversation.

Part II, we delve into the underlying hypothesis regarding the speaker, explore speech recognition, and delve into sentiment analysis. Section III elaborates on the proposed system, while Section IV outlines the characteristics of the experimental configuration. Following that, Section V showcases the results obtained and provides an extensive analysis. Finally, Episode VI concludes the project.

II. LITERATURE REVIEW AND CONTEXT

A. Sentiment Analysis

This method, also known as Sentiment Analysis SA, the objective is to ascertain whether a document conveys positive or negative sentiment through the analysis of sentiments expressed in textual form. Various techniques are employed in sentiment analysis research, including maximum entropy, decision trees, Naïve Bayesian, and support vector machines.

Mostafa et al. [4] classify sentences in each text as subjective or objective, subsequently processing the subjective portions using standard machine learning techniques to enable the polarity classifier to disregard false or irrelevant terms. Gathering and classifying data obtained from the analysis at the sentence level is time-consuming, and testing this process poses challenges. For hypothesis testing we use the following: Naive Bayes, Linear Support Vector Machines and VADER [6]. We compare these methods to determine the best-performing algorithm for our preferred use.

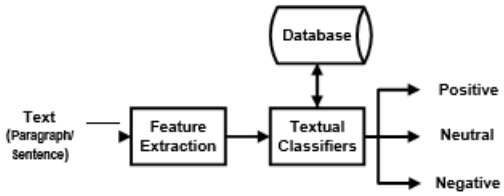


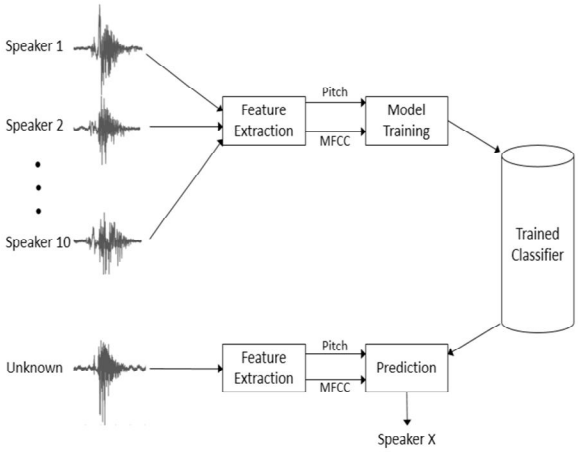
Fig. 1. Framework of Generic Sentiment Analysis System

B. Voice Recognition

Speech recognition is the ability of a machine, training model, or software to recognize words and sentences in human speech and convert them into a computer-readable format. The format can be used to perform other functions. In this work, we use speech recognition methods such as Google Speech Recognition, Bing Speech, and Sphinx4 [5]. Compare to choose the option that best suits your design goals.

C. Speaker Recognition

Speaker recognition involves identifying an individual based on their voice patterns and unique characteristics. This test has attracted great attention from researchers in the last eight years and is not genetic [7]. In speaker recognition, speaker-specific features are extracted from the speech signal, comprising various attributes capable of capturing emotional and semantic information unique to each speaker [8].



This work develops a speaker discriminant system using the Mel Frequency Cepstrum constant (MFCC). To find commonalities between the speech samples, the MFCCs for a variety of speakers' speech samples are extracted and compared with each other.

1) Feature Extraction

To achieve a higher accuracy rate, the unique speaker discriminant feature must be extracted. Because it serves as the input for the portion that follows, the precision of this part is essential.

MFCC or Mel Cepstrum Coefficient aims to simulate the human ear's perception of sound through mathematical modeling. This involves distinctive patterns ranging from noise to mel frequencies, usually spanning from 300Hz to 5KHz. Below 1 kHz, the Mel scale maintains linearity, transitioning to logarithmic above 1 kHz. Each individual MFCC speaker consistently records the intensity linked with each Mel frequency bin. This characteristic facilitates the recognition of American speakers.

2) Feature Matching

Dynamic time Wrapping (DTW)--- DTW is a rule that is understood as a technique of dynamic programming by Stan Republic of El Salvador et al. [7]. This rule is used to measure the similarity between two statistics that change corresponding to speed or time. If the series is "warped" non-linearly—that is, made tensile or shrunk on its time axis—just once, this system is also likely to detect the optimal alignment between the days' worth of data. Then, by using this warp between given two statistics, one can observe resemblance or complimenting areas between those two statistics. DTW's basic idea is to find two major dynamic patterns that possess similarities with one other and then imply a least possible distance amid them.

Ultimately, the statistic is encapsulated by employing various computational methods for measuring distance or similarity, such as geometric distance, Canberra Distance, and Correlation. A thorough comparison of these methods is presented in the results section of the paper.

III. PROPOSED SYSTEM

In our research, we are typically suggesting a model to analyze sentiments that makes the use of alternatives picked from the signals of speech provided to comprehend the emotion that speaker feels at the time when the phrase is spoken. There are four steps in the method: 1) VAD-equipped pre-treatment; 2) Speech Recognition; 3) Speaker Identification; and 4) Sentiment Analysis.

There is a system that detects the voice activity by receiving the signals. Later it uses these to identify and distinguish the phrase sounds from the signals received. These sounds are stored as chunks of information, which are subsequently transferred to the system for recognition of speech and speaker discrimination so as to identify the speaker's identity and content. The recognition system of speaker labels the segments with the IDs of speakers. It is relevant to mention that this system operates autonomously, identifying if the segments belong to the same speaker or two different speakers and labeling them as such. These segments are then transformed to text by the technology of speech recognition. In addition, the algorithm compares the transcribed text with the Speaker Id. It continues like a conversation inside the data. The textual output generated by a speaker-exclusive speech recognition system holds the potential to estimate the sentiment emphasized by the speaker at the moment of recording the statement. The whole method is expressed pictorially in the Figure 2.

IV. EXPERIMENTAL CONFIGURATION

A. Data set

Twenty-one audio files that were recorded in highly regulated environments make up our dataset [10]. Three distinct scripts are utilized for oral communication between two individuals. In these recordings, seven speakers—four men and three women—are fully engaged. The chats were labeled as per the current scenario. The audio has been recorded in form of mono-tracks for an approximation of ten seconds, with a 16 KHz sample rate.

A sample of dataset is displayed in Figure 3.

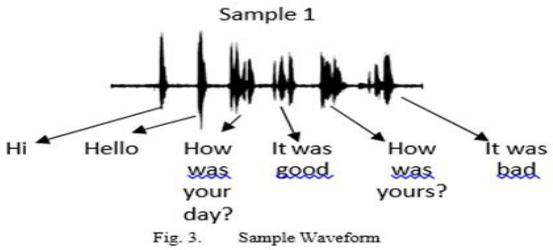


Fig. 3. Sample Waveform

B. Tests and Measurements

The methodology of speech recognition, speaker recognition, and sentiment analysis are all used in our proposed system. We have provided an explanatory analysis of the tests that have been carried out using various instruments and methods. The voice recognition tools that are utilized are Sphinx4, Bing voice API, and Google Speech API. WWR was also employed as a performance metric. We often employ MFCC as the feature for speaker recognition and DTW with a variety of distance computation methodologies, such as national capital, geometer, and correlation that have been used for matching of features. Additionally, recognition rate has been used as a statistic for measure of performance. Normal sentiment analysis datasets, such as the Twitter and Product Review databases [6], have been utilized gauge the accuracy of system in the process of sentiment analysis.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

V. RESULTS

A. Results pertaining to the Automatic Speech Recognition Engine

First, using entirely distinct speech recognition software, the audio recordings from the dataset were transformed to the text files. The WRR acquired for multiple scripts delivered by absolutely different speakers is displayed in Table 1. Like how money supply indicates the first male speaker, F1 indicates the first female speaker. Share values are used to express the WRR.

TABLE I. WRR for Sphinx4:

Speaker1	Speaker2	Script1	Script2	Script3
M1	M2	46.67	23.08	62.50
M2	M3	33.33	30.77	18.75
M3	M1	26.67	23.08	25.00
M2	M4	53.33	30.77	56.25
F1	F2	46.67	38.46	31.25
F2	F3	26.67	30.77	37.50
F3	F1	33.33	38.46	25.00

The WRR acquired by utilizing the Google Speech API to decipher the signals obtained from speech is tabulated in Table 2. Constant datasets, or scripts, are used, and the same people are frequently used to compare the outcomes. This can be carried out to verify the tools on an equal footing.

TABLE II. WRR for Google Speech API:

Speaker 1	Speaker 2	Script1	Script2	Script3
M1	M2	93.33	84.62	81.25
M2	M3	86.67	92.31	75.00
M3	M1	86.67	84.62	43.75
M2	M4	80.00	76.92	68.75
F1	F2	86.67	84.62	81.25
F2	F3	93.33	84.62	37.50
F3	F1	80.00	92.31	75.00

Similarly, Table 3, has WRR obtained for the same dataset but by using Bing Speech API.

TABLE III. WRR for Bing Speech:

Speaker 1	Speaker 2	Script1	Script2	Script3
M1	M2	100.00	92.31	87.50
M2	M3	93.33	84.62	87.50
M3	M1	86.67	92.31	93.75
M2	M4	86.67	84.62	81.25
F1	F2	80.00	84.62	93.75
F2	F3	93.33	92.31	87.50
F3	F1	86.67	76.92	93.75

The average of the WRR obtained from the previous table is given in Table 4.

TABLE IV. Average:

Speech Engine	Average for Script1	Average for Script2	Average for Script3	Average
Sphinx4	38.10	30.77	36.61	35.16
Google Speech API	86.67	85.71	66.07	79.48
Bing Speech API	89.52	86.81	89.29	88.54

B. Results for Speaker Discrimination System:

Figure 5 shows how accurate talker identification is when there are multiple possibilities available. There were somewhere between one and twenty-six possibilities. Because we used the feature mapping technique in our study, i.e. dynamic time wrapping. Differrent methods of calculating distance, such as geometers and Australian capital square measures with DTW and comparisons. Graph 3 displays the comparison of accuracy and variety of alternatives of graph. After using some twelve or fourteen likelihoods, it was found that the system is quite accurate; as a result, we typically take thirteen options to navigate the system.

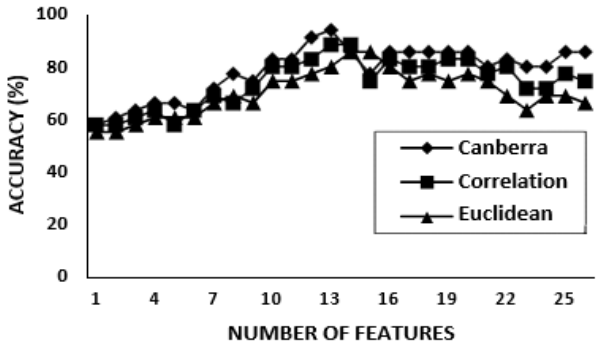


Fig. 6. Accuracy vs Number of Features

C. Results for Sentiment Analysis System:

Table 5 displays the accuracy of several algorithms utilized in this model for sentiment analysis, including Naive Bayes, Linear SVM, and, VADER.

TABLE V. Accuracy Sentiment

Method	Twitter Dataset	Movie Review
Naive Bayes	84	72.8
Linear SVM	88	86.4
VADER	95.2	96

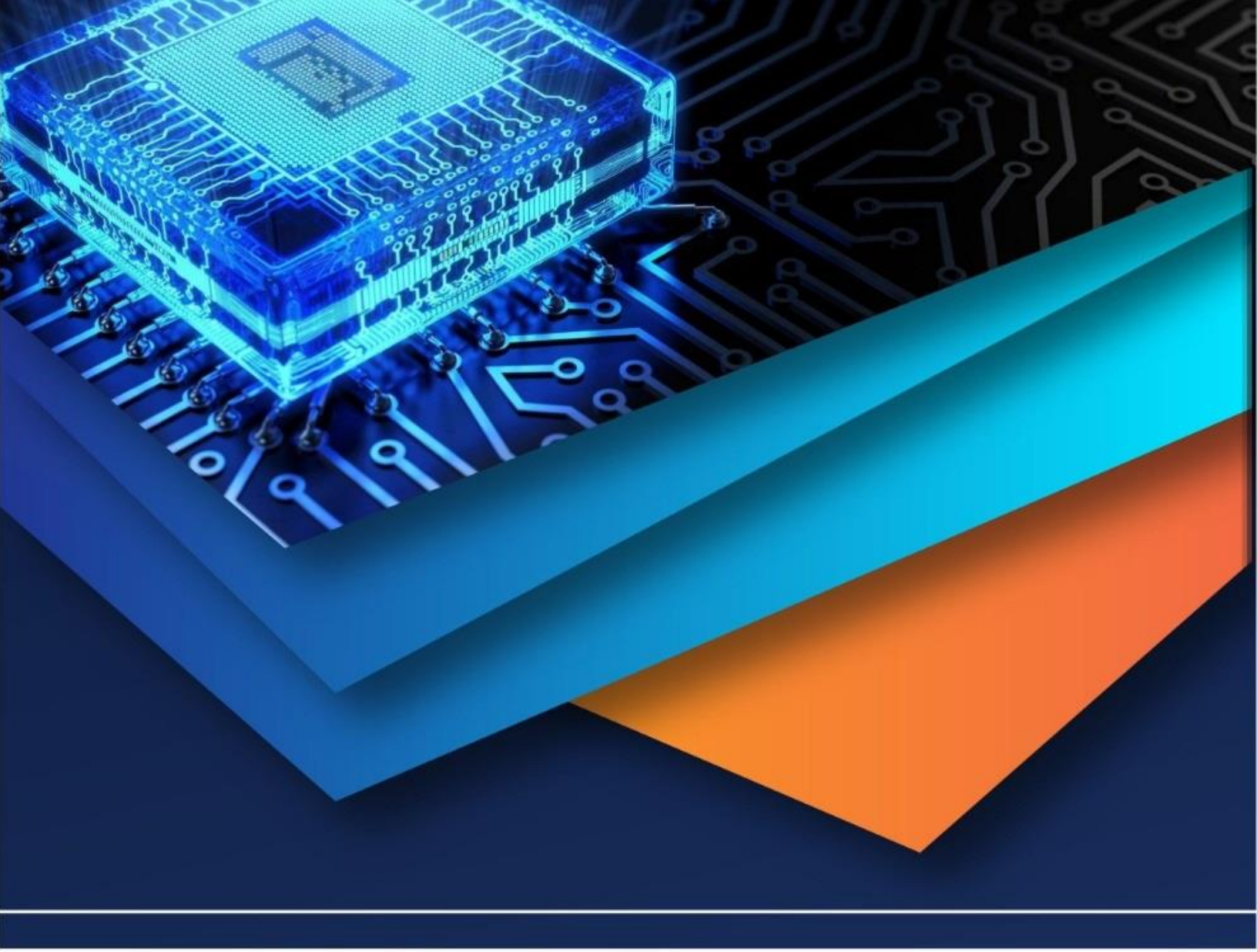
VI. CONCLUSION AND FUTURE WORK

This research paper coins a generalized model that takes in use the audio file from two speakers as an input, mechanically converts the audio to text, later, plays back speaker identification so as to study the identities of the speaker and the content provided. We have designed an easy to use system that attempts to do the precursory task in the duration of this study. The system functions effectively with artificially generated datasets; however, we tend to perform better when grouping larger datasets to improve the system's measurability. Even though the system accurately interprets the sentiments of the speakers in casual dialogue, it has certain shortcomings. Firstly, it is unable to distinguish between two speakers speaking at the same time. Only one speaker should speak at a time during a speech. In the future, we would seek to solve these issues and enhance the system's measurability and accuracy.

REFERENCES

[1] Pang, B., & Lee, L. (2004, July). Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.
[2] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on associationfor computational linguistics (pp. 115-124). Association for Computational Linguistics.
[3] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Associationfor Computational Linguistics.
[4] Shaikh, M., Prendinger, H., & Mitsuru, I. (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. Affective Computing and Intelligent Interaction, 191- 202.

[5] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., ... & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.
[6] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification.
[7] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis.
[8] Dos Santos, C., & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts.
[9] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the conference on empirical methods in natural language processing (EMNLP), 1631-1642.
[10] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
[11] Xia, R., & Zong, C. (2011). Ensemble of feature sets and classification algorithms for sentiment classification.
[12] Zhang, Y., & Wallace, B. C. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.
[13] Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network.
[14] Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., & Wang, F. (2012). Sentiment Analysis on Twitter with Semi-supervised Learning.
[15] Severyn, A., & Moschitti, A. (2015). Twitter Sentiment Analysis with Deep Convolutional Neural Networks.
[16] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
[17] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media
[18] Towards Data Science - A Gentle Introduction to Sentiment Analysis
[19] Analytics Vidhya - Text Mining and Sentiment Analysis: A Beginner's Guide
[20] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pretraining.
[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need.
[22] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Understanding.
[23] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J. (2019). Hugging Face's Transformers: State-of-the-art Natural Language Processing
[24] Dathathri, R., Madotto, A., Lan, Z., Ni, J., Dong, L., & Gao, J. (2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation.
[25] Shao, L., Feng, X., Wallace, B., & Bengio, Y. (2017). Generating high-quality and informative conversation responses with sequence-to-sequence models
[26] Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., & Bengio, Y. (2016). A hierarchical latent variable encoder-decoder model for generating dialogues.
[27] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
[28] Brown, T. B., Mané, D., Roy, A., Abolafia, D. A., & others. (2020). Language Models are Few-Shot Learners.
[29] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality.
[30] Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). Deep Reinforcement Learning for Dialogue Generation.
[31] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues.
[32] Zhang, S., & Xu, W. (2017). Neural Network-based Abstract Generation for Opinions and Arguments.
[33] Kannan, A., Kurach, K., Ravi, S., & Bengio, Y. (2016). Adversarial Training for Text-to-Image Synthesis.
[34] Zhang, Z., & Lapata, M. (2014). Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY