ASSAMESE DIALECT CHARACTER LEVEL IDENTIFICATION USING CONVOLUTION NEURAL NETWORK

Abstract

This paper introduces an approach for automatically identifying four Assamese dialects: Central dialect, Eastern dialect, Kamrupia dialect and Goalporia dialect. The suggested method involves utilizing a character-level convolution neural network along with dialect embedding vectors. These embedding vectors are compact representations derived from linguistic attributes and are employed to differentiate among the four dialects. We presented three models that share a consistent architecture, differing only in the initial layer. The convolution layer's input in the initial system is a character representation encoded in a one-hot manner. Prior to the convolution layer, the embedding layer is introduced in the second system. The third system employs a recurrent layer prior to the convolution layer. The most favorable outcomes were achieved by employing the third model, which attained a F1-score of 74.69%.

Keywords: CNN; ADI; dialect; embedding vectors; Assamese

Author

Hem Chandra Das
Department of Computer Science & Technology
Bodoland University
Kokrajhar, Assam, India
hemchandradas78@gmail.com

I. INTRODUCTION

Dialects encompass unique speech patterns within a language, used by individual's in particular geographical areas. These variations encompass differences in language structure, pronunciation, and intonation. The pronunciation of words is shaped by factors including social standing, cultural heritage, location, and education [1]. Advanced technologies capable of categorizing and distinguishing dialects hold promise in improving the functionality of interactive speech applications. The features specific to dialects significantly influence the performance of automated speech recognition (ASR) and human-computer interface (HCI) systems [2]. Incorporating understanding of dialects into pronunciation dictionaries and acoustic training can significantly enhance the effectiveness of speech-based systems [3]. Within forensic science, dialect recognition is employed for tasks like verifying speakers, validating speech, and creating speaker profiles [4]. The identification of dialects can also enhance the interaction between users and machines. Additionally, dialect recognition systems have practical uses in dialogue processing, retrieving spoken documents, translating spoken language, and ensuring accurate speech-to-text conversion [5]. Furthermore, the capability for dialect recognition is valuable in pinpointing native languages and has applications across domains like medicine, archiving and retrieving previously spoken content, the media sector, and beyond [6].

Assam, situated in north-eastern India, hosts the Assamese language, which is utilized by the local populace. Within Assam, a range of regional Assamese dialects exists, differing in terms of pronunciation, sentence structure, and vocabulary. These variations encompass the Central dialect, found in and around Nagaon district; the Eastern dialect, spoken in Sibsagar and adjacent districts; the Kamrupi dialect, used in Kamrup, Nalbari, Barpeta, Kokrajhar, and parts of Bongaigaon district; and the Goalporia dialect, spoken in Gopalpara, Dhuburi, and a section of Bongaigaon district. Recognizing these dialects is pivotal for the development of an inclusive Assamese voice recognition system capable of accurately identifying words spoken in Assamese and its diverse dialects.

Assamese, deriving its roots from the Indo-Aryan language family, holds the position of the official language in Assam and is extensively spoken in the northeastern region of India. Local speakers articulate it as "axamija." Originating from Sanskrit, Assamese, which has evolved over an extensive historical period, absorbed vocabulary from non-Aryan languages [7]. Beyond the inherent phonetic diversity, Assamese showcases various regional dialects found across the state. The languages spoken in upper and lower Assam can be broadly divided into two main groups. The colloquial Assamese standard is rooted in upper Assam. Linguist Banikanta Kakati categorized Eastern and Western Assamese based on linguistic resemblances. Western Assamese refers to the language spoken between undivided Kamrup and Goalpara, whereas Eastern Assamese encompasses the area from Sadiya to Guwahati. Further variations within Western Assamese among speakers from Goalpara and Kamrup have led to the emergence of sub-dialects. Linguist G. C. Goswami later introduced the central dialect, situated between upper and lower Assamese. He classified regional dialects into Upper Assamese, Lower Assamese, and Central Assamese, with Upper Assamese extending from Nagaon to Sonitpur, Lower Assamese spanning from east Kamrup to Goalpara, and Central Assamese covering Darang to Morigaon and Kamrup [8-9].

Recent investigations conducted by contemporary linguists have categorized the Assamese dialects into four primary groups: Eastern, Central, Kamrupi, and Goalpariya [10]. These studies illustrated the intra-divisional portrayal of each dialect, as illustrated in Figure 1. The Kamrupi dialect, spoken in districts such as Barpeta, Nalbari, and Kamrup, exhibits sub-dialects known as Barpetiya, Nalbariya, Kamrupi, and South Kamrupi. A notable trait of Kamrupi is its use of initial stress, resulting in word shortening compared to the penultimate stress observed in Eastern dialects. For instance, the term "vegetable gourd" is pronounced as "/kumra/" in Kamrupi dialect, whereas in standard Assamese, it's "/komora/." Additionally, Kamrupi introduces more high vowels when compared to Eastern Assamese, where medial vowels predominate. The Goalpariya dialect is spoken in the Goalpara district, encompassing contemporary Dhubri, Goalpara, Kokrajhar, and Bongaigaon districts. It features Eastern and Western Goalpariya sub-dialects used in Goalpara, Bongaigaon, and Dhubri regions. The Goalpariya dialect shares morphological and phonological characteristics with Bengali, another prominent language in India [11]. Variations in Assamese dialects often arise from dissimilarities in the pronunciation of vowel sounds. As an example, the term "king" is articulated as "/rɔza/(ৰজা)" in standard Assamese, whereas in Kamrupi and Goalpariya dialects, it's "/raza/(ৰাজা)" [12]. Consequently, the Assamese language demonstrates notable disparities across regions and speakers, primarily due to these variations in vowel usage. We employed a technique involving a Convolution Neural Network at the character level to detect Assamese dialects by utilizing lexical and dialect embedding attributes.

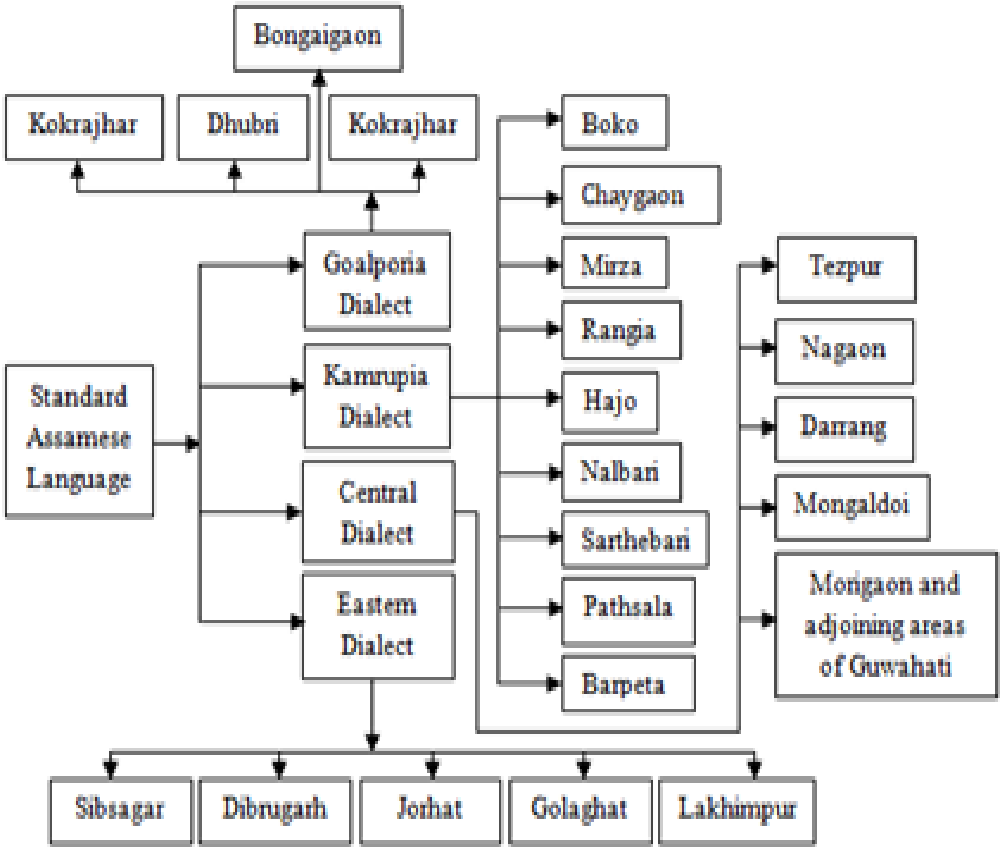

Figure 1: Assamese Dialect's Internal Division.

II. LITERATURE REVIEW

A suggested approach for identifying Assamese dialects involves using a fuzzy neural network-based system that leverages formants and prosodic information extracted from vowel sounds [13]. Shashidhar G. Koolagudi [14] developed a system capable of recognizing fifteen different Indian languages, such as Assamese, Bengali, Hindi, and English, using twenty-one Mel Frequency Cepstral Coefficient (MFCC) features obtained from audio signals. Another study introduced an automatic language recognition system that utilized K-means clustering on MFCCs and employed Support Vector Machines (SVM) for classification, focusing on three Indian languages: English, Hindi, and Tibetan [15]. In terms of dialect identification, Ismail et al. [16] conducted research on Kamrupi, Goalporia, and Assamese dialects, employing GMM and GMM-UBM (Gaussian Mixture Model with Universal Background Model) methods. GMM-UBM yielded higher identification rates compared to GMM alone. Sarma et al. [13] utilized a Feed Forward Neural Network (FFNN) to categorize dialects based on vowel sounds, demonstrating its effectiveness with improved classification rates. Das and Bhattacharjee [17] introduced a technique that utilized Gaussian Mixture Model (GMM) and GMM-UBM for identifying Assamese dialects, achieving a remarkable identification accuracy of 97.57%. Additionally, Sarmah and Dihingia [18] employed a random forest approach to distinguish Assamese dialects based on the acoustic properties of Assamese vowels, resulting in a classification accuracy of 94.0%.

Conversely, additional investigations were conducted in the realm of identifying dialects within written text. In 2013, Elfardy and Diab employed labels at the word level to extract features at the sentence level. Subsequently, these features were utilized to assign the appropriate dialect label to the sentence [19].

Zaidan and Callison-Burch [20] utilized a labeled dataset named Arabic Online Commentary (AOC) to develop a system for identifying the Arabic dialect sentences. Following suit, the same dataset was also used to train a linear support vector machine classifier with binary characteristics based on words [21]. Their objective was to distinguish between Egyptian dialect and Modern Standard Arabic (MSA). In a similar context, the AOC corpus exhibited a degree of uniformity, given its origins from specific sources. Because of this homogeneity, it was difficult for models developed using this dataset to transfer their knowledge to new domains. Furthermore, character-based n-grams were more effective in differentiating between Egyptian dialect and MSA than word-based n-grams [22] .

Lexical variables obtained from a speech recognition system were combined with acoustic qualities to create a more reliable classifier than classifiers that only use acoustic or lexical features [23]. The Discriminating between Similar Languages (DSL) shared task's Automatic Dialect Identification (ADI) components first appeared in 2016. The dataset used for this specific subtask consisted of transcribed speech samples in Modern Standard Arabic (MSA), alongside four distinct dialects: Levantine (LAV), Gulf (GLF), Egyptian (EGY), and North African (NOR) [23]. The majority of participants in this particular task employed character n-grams, and the most impressive outcome was attained by utilizing the SVM (Support Vector Machine) classifier on character n-grams ranging from 1 to 7 [24]. A character-level convolutional neural network was utilized, employing a similar method to what we are currently using [25]. However, their focus was solely on using ASR transcripts of Arabic speech due to the unavailability of acoustic features during that period. The ADI

Shared Task, the dataset comprised both the initial audio files and a set of basic audio attributes referred to as i-vectors. These were accompanied by ASR transcripts derived from Arabic speech obtained within the Broadcast News field. The Kernel Discriminant Analysis (KDA) classifier, which was trained on both character n-grams and the i-vector characteristics using a variety of kernel functions, was used to achieve the task's most significant result [26].

III. EXPERIMENT SETUP

1. **CNN at Character-Level :** Initially developed for image processing, Convolutional Neural Networks (CNN) have demonstrated remarkable performance in the area of computer vision [27][28][29]. Subsequently, their application has extended to NLP (Natural Language Processing) tasks, surpassing conventional models like words in a bag, n-grams, and their variations based on Term Frequency-Inverse Document Frequency (TFIDF)[30][31]. Figure 2 illustrated, the outlined architecture outlines the character-level CNN model harnessed for identifying Arabic dialects. The assignment is framed as a multi-class classification problem using our methodology. Our goal is to predict r using s and t given an ASR transcript labelled as $t^{(i)}$, 600-dimensional feature vectors for dialect embedding indicated as $s^{(i)}$, and their corresponding label $r^{(i)}$. We constructed a neural network classification algorithm that takes both the transcript, encoded as a one-hot array of characters (adjusted to a predetermined maximum length through padding or truncation), and the matching feature vector for dialect embedding as inputs. As depicted in Figure 1, the transcript text undergoes convolutional processing before reaching a softmax layer, the embedding vector, however, is sent straight into a different softmax layer. The average of these two softmax layers makes up the final network output, which results in a distribution of probability across the 4 Assamese dialects. The network architecture comprises the subsequent layers:

   - **Input:** The input layer converts each character into a corresponding one-hot vector representation.

   - **Recurrent:** This layer is employed either for embedding purposes or for incorporating a recurrent GRU layer, aiming to grasp the contextual details about the character [32].

   - **Convolutional:** This layer contains a variety of feature maps and filter widths that are utilised for character windows to create new traits. Each convolution is followed by a batch-normalization layer and a Rectified Linear Unit (ReLU) nonlinearity layer[33][34].

   - **Max-Pooling:** Each filter's feature map should be subjected to the max-over-time pooling technique, and this filter should use the greatest value as a feature [35]. To avoid over-fitting, a dropout layer is applied after the max-pooling operation [36].

   - **Softmax:** One softmax layer is used to manage lexical attributes, and the other is used to manage embedding features.

- **Output:** The mean of the outputs from the two softmax layers serves as the final output, which depicts the the likelihood distribution across the labels.

Based on the outcomes of cross-validation, we employed the subsequent variables for configuring the architecture of a neural network:

- Length of the sentences is limited to a maximum of 256 characters.
- The embedding length dimension is 128.
- GRU layer is set to 128 units.
- Convolutional filter size ranges from 2 to 8..
- 256 convolution filter feature map is consider for each filter.
- Dropout is applied with drop probability 0.2.

2. **Database:** The examination of existing literature highlights the absence of a standardized database for Assamese language and its various dialects. To address this gap, a novel database was constructed, comprising speech samples representing different dialect groups. This database, employed in this study, was curated by collecting spoken samples from native speakers belonging to four distinct Assamese dialects, involving scripted speech. The recording sessions primarily engaged individuals from rural areas who were either native to the region or had resided there for an extended period. Most speakers possessed minimal formal education, typically up to or below matriculation level, which indicated limited exposure to written and standard Assamese and thus preserved a more authentic native dialect. Around 90% of the selected speakers exclusively spoke Assamese. The age range of the speakers spanned from 25 to 65 years. Recordings were conducted using a Sony voice recorder at a 44.1 kHz sampling rate and 16-bit mono resolution per sample, maintaining a relatively quiet recording environment.

The speech dataset consisted of 10 individuals, evenly split between genders, and capable of reading and speaking, representing the entirety of Assamese dialects. There were the same amounts of speakers in each dialect region. Roughly 3 minutes of audio recordings were collected from 10 distinct speakers within each dialect, resulting in approximately 30 minutes of voice data for each of the four dialects. A phonetically diverse script was prepared for recording speech samples, consistently used across all dialects. To ensure quality, subjective listening tests were conducted by individuals who were not involved in the recording process and are from the relevant dialect groups.

The dataset consists of four feature sets with a total of 1,677 utterances for validation and 15,612 utterances for training. The Siamese neural network methodology produces 600-dimensional dialect embeddings for every utterance from linguistic characteristics [37].

We only used the dialect embedding characteristics and ASR transcripts during our studies.

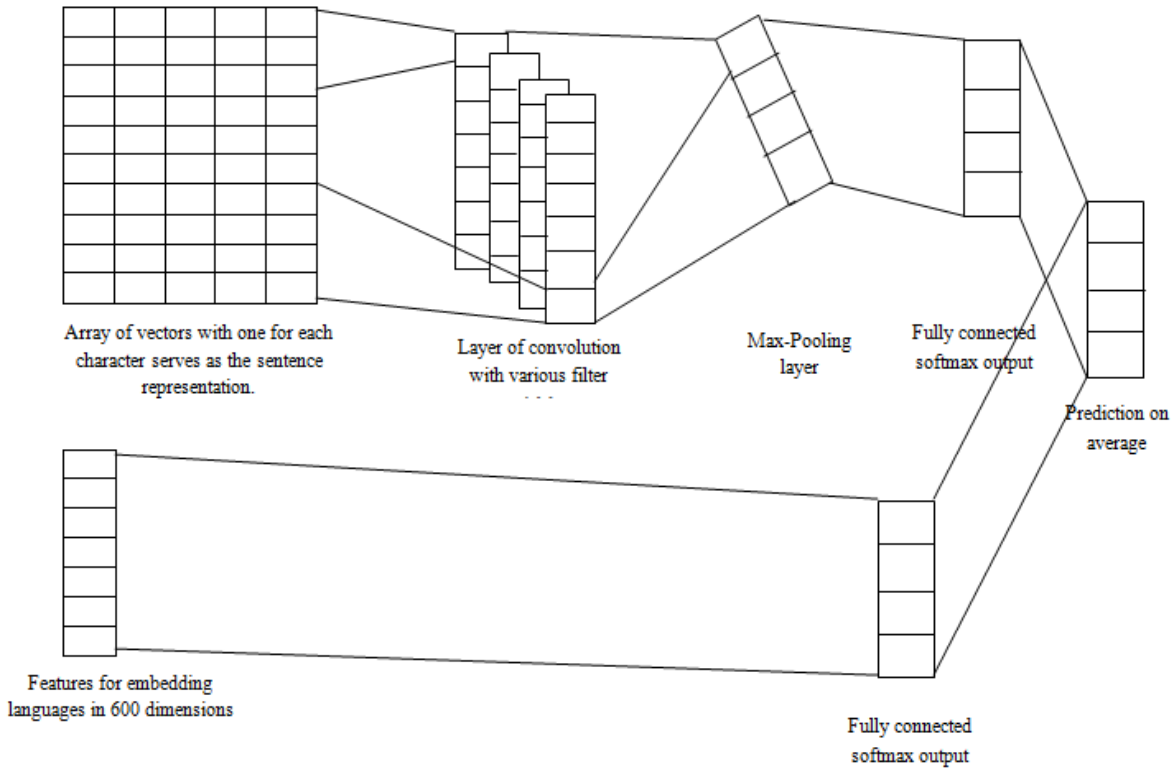

Figure 2: Architecture for CNN at the Character Level.

IV. RESULTS AND DISCUSSION

1. **Result of Cross-Validation :**We used cross-validation with 10-fold approach after merging the validation data from the shared task with the training data. We assessed three different configurations and compared them with a classifier focused on TF-IDF characteristics, which was a LR (Logistic Regression) classifier created with the scikit-learn package [38]. The outcomes are detailed in Table 1.

Table 1: Results of Cross-Validation

System	Accuracy
Using TF-IDF features in logistic regression	0.7825
One-hot input encoding for CNN	0.7932
Embedding layer for CNN	0.7913
GRU recurrent layer on CNN	0.8075

2. **Test Set Results :** Table 2 displays the outcomes of three distinct executions. Except for the input used by the convolution layer, we maintained the same setup during all three runs. In the initial execution, the convolution layer received direct input from the character sequences' one-hot encoded vectors. The convolution layer was preceded by an embedding layer in the second execution and by a GRU recurrent layer in the third execution.

Findings showed that using a recurrent layer produced better results than using a traditional embedding layer or feeding the convolution layer directly with the one-hot encoded representation. The recurrent layer network needed to train for almost five times as long as the network without them, thus this improvement came at a significant time expense. In the evaluation of the shared GDI job, the F1-weighted scores of the submitted systems were used to rank them.. The result achieved with an F1-weighted score of 74.49%. Figure 3 portrays the confusion matrix for most successful run. The matrix highlights that the Central dialect posed the most confusion, frequently being misidentified as the Eastern dialect.

Table 2: Results of Three Runs, Using Bold for the Best Run

System	Accuracy
Baseline in Random	0.3521
One-hot input encoding for CNN	0.7223
Embedding layer for CNN	0.7171
GRU recurrent layer on CNN	0.7449

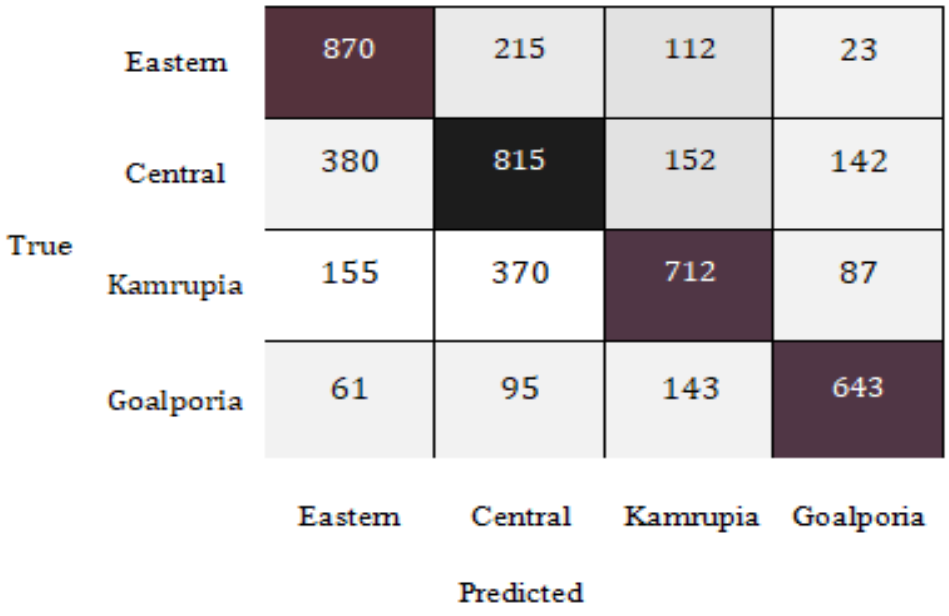

Figure 3: Confusion Matrix GRU Recurrent Layer on CNN.

V. CONCLUSIONS

This study involves extracting features using a character-level CNN from textual data, coupled with dialect embedding characteristics derived from the same text. In the most successful of our submissions, a GRU recurrent layer was deployed as an embedding layer prior to the convolutional layer. Nevertheless, the benefit gained from incorporating the recurrent layer was marginal when weighed against the extensive training time required for a network equipped with such a layer, in contrast to a network with a conventional embedding layer.

REFERENCES

[1] J.K. Chambers and P. Trudgill, Dialectology, Cambridge :Cambridge University Press,1992.
[2] E. Ferragne and F Pellegrino, "Automatic dialect identification: A study of British English," Speaker Classification II: Selected Projects, vol. 4441, pp.243-257, 2007.
[3] M. Najafian, A. DeMarco, and S. Cox, "Unsupervised model selection for recognition of regional accented speech," in Fifteenth annual conference of the international speech communication association, 2014.
[4] M.J. Harris, S.T. Gries, and V.G. Miglio, "Prosody and its application to forensic linguistics," Linguistic evidence in security, law and intelligence, vol.2, pp.11-29,2014.
[5] H. Li, B. Ma, and K.A. Lee, "Spoken language recognition: from fundamentals to practice," Proceedings of the IEEE, vol. 101, pp.1136-1159, 2013.
[6] S. Gray and J.H.L. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system," in IEEE Workshop on Automatic Speech Recognition and Understanding, 2005, pp.35-40.
[7] G. C. Goswami, Structure of Assamese. Gauhati University, India: Dept. of publication, 1982.
[8] B. Bharali, Kamrupi Upabhasha: eti adhyayan. Guwahati, Assam, India: Banlata, 2008.
[9] U. Goswami, A study on Kamrupi: a dialect of Assamese, Assam. India: Dept. of Historical Antiquarian Studies, 1970.
[10] Resource Centre for Indian Language Tech-nology Solutions, Indian Institute of Technology, Guwahati. https://egovindia.wordpress.com/2006/06/21/resource-centre-for-indian-language-technology-solutions-rcilts-iit-guwahati/ assamese-language.pdf.
[11] B. Bharali, K. Talukdar, Goalpariya Upabhasha: Rup Boichitrya. Kumarpara, Guwahati, Assam, India: Shib Prakashan, 2012.
[12] M. Sarma and K.K. Sarma, Phoneme-based speech segmentation using hybrid soft computing frame-work. New Delhi: Springer, 2014.
[13] M. Sarma and K.K. Sarma, "Dialect Identification from Assamese speech using prosodic features and a neuro fuzzy classifier," in Proceedings 3rd International Conf. on Signal Processing and Integrated Net-works (SPIN), Feb. 2016, pp. 127-132.
[14] S.G. Koolagudi, D. Rastogi, and R.S. Rao, "Identifi-cation of language using mel-frequency cepstral coefficients (MFCC)," Procedia Engineering, vol. 38, pp.3391-3398, 2012.
[15] V.K. Verma and N. Khanna, "Indian language identification using k-means clustering and support vector machine (SVM)," in Proceedings 2013 Students Conf. on Engineering and Systems (SCES), Apr. 2013, pp.1-5.
[16] T. Ismail and L.J. Singh, "Dialect identification of Assamese language using spectral features," Indian Journal of Science and Technology, vol.10, pp.1-7, 2017.
[17] H.C. Das and U. Bhattacharjee, "Identification of Four Major Dialects of Assamese Language Using GMM with UBM," In Lecture Notes in Electrical Engineering 888, Gupta, D., Goswami, R.S., Banerjee, S., Tanveer, M., Pachori, R.B. (eds), Springer Nature, 2021, pp. 311-319.
[18] P. Sarmah and L. Dihingia, "Assamese Dialect Identification from Vowel Acoustics," in Proceedings data Engineering for smart systems: Proc. of SSIC 2021, Nov. 2022, pp. 313-322.
[19] H. Elfardy and M. Diab, "Sentence level dialect identification in Arabic," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol.2, 2013, pp.456-461.
[20] O.F. Zaidan and C. Callison-Burch, "Arabic dialect identification," Computational Linguistics, vol.40, pp.171-202, 2014.

[21] C. Tillmann, S. Mansour, and Y. Al-Onaizan, "Improved sentence-level Arabic dialect classification," in Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects,2014, pp.110-119.
[22] K. Darwish, H. Sajjad and H. Mubarak, "Verifiably effective arabic dialect identification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1465-1468.
[23] A. Ali et al. "
[24] 2013.
[25] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Automatic dialect detection in arabic broadcast speech," arXiv preprint arXiv:1509.06928,pp. 2934–2938, 2015.
[26] Ç. Çöltekin and T. Rama, "Discriminating similar languages with linear SVMs and neural networks," in Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016, pp.15-24.
[27] Y. Belinkov and J. Glass, "A character-level convolutional neural network for distinguishing similar languages and dialects," arXiv preprint arXiv:1609.07568, pp.145-152, 2016.
[28] R.T. Ionescu and A. Butnaru, "Learning to identify arabic and german dialects using multiple kernels," in Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), 2017, pp.200-209.
[29] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol.60,pp.84-90, 2017.
[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, transactions on pattern analysis and machine intelligence, vol.35, pp.221-231, 2012.
[31] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning, 2008, pp.160-167.
[32] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," Advances in neural information processing systems, vol.28, pp. 649–657, 2015.
[33] J. Chung, C. Gulcehre, K.H. Cho and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555,2014.
[34] X. Glorot, A. Bordes and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the fourteenth international conference on artificial intelligence and statistics,2011, pp.315-323.
[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift,"in proceedings of International conference on machine learning, 2015, pp.448-456.
[36] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural language processing (almost) from scratch," Journal of machine learning research, vol.12, pp. 2493-2537, 2011.
[37] N. Srivastava, G. Hinton and A. Krizhevsky, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol.15, pp. 1929-1958, 2014.
[38] S. Shon, A. Ali and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," arXiv preprint arXiv:1803.04567, pp. 98–104, 2018.
[39] F. Pedregosa et al. "Scikit-learn: Machine learning in Python," The Journal of machine Learning research, vol.12, pp. 2825-2830,2011.

