Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 08, August - 2025

Named Entity Recognition for Konkani Speech

Shalaka Naik Dessai
Department of Information Technology
Goa Engineering College
Ponda, Goa

Abstract: **Named Entity Recognition (NER) is the procedure of recognizing and classifying all the proper nouns into pre-defined classes such as locations, persons, organization, date and others. NER has many uses such as Speed up the hiring process by summarizing applicants' CVs, improve internal workflows by categorizing employee complaints and questions, Enable students and researchers to find relevant material faster by summarizing papers and archive material and highlighting key terms, topics, and themes and has many other uses. To work on NER in Indian languages is a difficult and challenging task and also limited due to scarcity of resources, but it has started to appear recently. A named entity recognizer (NER), an essential tool for natural language processing (NLP), is presented for the first time for the Konkani speech as per my knowledge. In this paper NER for Konkani Speech is proposed. The Proposed model is able to identify named entities from Konkani audios and classify them into pre-defined categories like person, location, date and organization.**

***Keywords*-Named Entity Recognition, Convolutional Neural Network, Support Vector Machine, Audio Classification.**

1.INTRODUCTION:

According to the 2001 Census, around 122 important Indian languages and 1599 different languages are spoken in India. The on-line usage of Indian languages has increased, however they still generally tend to be afflicted by numerous problems inclusive of insufficient online processing support, incapacity to deal with numerous scripts, and shortage of study work performed because of inadequate funding in those areas. Konkani is the authentic language of the country of Goa. About 2.5 million people, 0.2% of the full populace of India, communicate Konkani, that is a completely small quantity in comparison to audio system of the important Indian languages. Since there may be rarely any studies paintings performed on Natural Language Processing for the Konkani language, it turns into more and more more hard to increase a Named entity recognition tool for Konkani speech with out the necessary supporting language processing tools.

Named entity recognition (NER) — sometimes referred to as entity chunking, extraction, or identification — is the task of identifying and categorizing key information (entities) in text. An entity can be any word or series of words that consistently refers to the same thing. Every detected entity is classified into a predetermined category. For example, an NER machine learning (ML) model might detect the word "super.AI" in a text and classify it as a "Company".

NER is a form of natural language processing (NLP), a subfield of artificial intelligence. NLP is concerned with computers processing and analyzing natural language, i.e., any language that has developed naturally, rather than artificially, such as with computer coding languages.

2.LITERATURE SURVEY:

The authors in the paper [1] presented a first study of end-to-end approach that directly extracts named entities from speech, though a unique neural architecture. Experimental results show that this end-to-end approach provides better results with F-measure=0.69 on test data than a classical pipeline approach to detect named entity categories with F-measure=0.65.

In paper [2] the authors have explored a variety of approaches for using external data to improve both pipeline and E2E approaches for spoken NER. They noted that E2E approaches are better able to take advantage of the external data by distilling information from the more semantically mature pre-trained text representations. On the other hand, pipeline approaches show minimal improvements from the use of external data. Their analysis also hints at E2E model's capability to focus on more entity-specific words despite being poorer at speech-to-text conversion than pipeline models.

The authors in paper [3] presented the first Vietnamese speech dataset for NER task and the first pre-trained public large-scale monolingual language model for Vietnamese They also proposed a new pipeline for the NER task from speech that overcomes the text formatting problem by introducing a text capitalization and punctuation recovery model (CaPu) into the pipeline. The model takes input text from an ASR system and performs two tasks at the same time, producing proper text formatting that helps to improve NER performance.

The authors in paper [4], developed a system that incorporates word, character, and morph representations which achieves competitive results on the Digitoday dataset. They concluded that transferring tags from the Estonian language using multilingual embeddings significantly improved the results on the out-of-domain Wikipedia test set. They also evaluated their system on two ASR output datasets, where one of them did not have capitalization and punctuation, which caused difficulties for their system, so in order to mitigate those difficulties, they converted the training set to lowercase and removed the punctuation in order to simulate ASR setting, which yielded significant improvement.

In paper [5] authors developed an end to end neural model based on bidirectional Long Short Term Memory (Bi-LSTM) for Hindi NER .They designed their model in two stages. In first stage, they used the unlabelled corpus to learn word embedding based on skip gram approach and glove approach. In second stage, their system used bidirectional LSTM. System’s embedding layers were initialized with learned word vectors for every word and then, system was trained end-to-end on labelled data.

Authors in paper [6] presented a learning-based named entity recognizer for Bengali that does not rely on manually constructed gazetteers in which they developed two architectures for the NER system. The corpus consisting of 77942 words is tagged with one of 26 tags in the tag set defined by IIT Hyderabad where they used CRF++ to train the POS tagging model.

Authors in paper[7] developed a domain-specific Tamil NER for tourism by using CRF. It handles morphological inflection and nested tagging of named entities with a hierarchical tag set consisting of 106 tags. A corpus of 94k is manually tagged for POS, NP chunking, and NE annotations. The corpus is divided into training data and the test data where CRF is trained with the former one and CRF models for each of the levels in the hierarchy are obtained. The system comes out with an F-measure of 80.44%.

In paper [8] authors used part of LERC-UoH Telugu corpus where CRF based Noun Tagger built with 13,425 words manually marked data and tested in a test set of 6,223 words and came out with F-average 91.95%. Then they create a NER based on the law a program with 72,152 words including 6,268 Named Businesses where they identified specific issues related to Telegu. The NER also later developed a CRF-based NER system for Telegu and obtain a total F rating of between 80% and 97% in various tests.

3 .DATA EXPLORATION AND ANALYSIS

A. Dataset

Audios used for this project is collected by recording voices of different speakers.

All the audios are stored in the WAV format as it helps as this format recreates the recording accurately without losing audio quality due to the format itself. The dataset contains 100 audios divided into 4 different classes:

- Person
- Location
- Organization
- Date

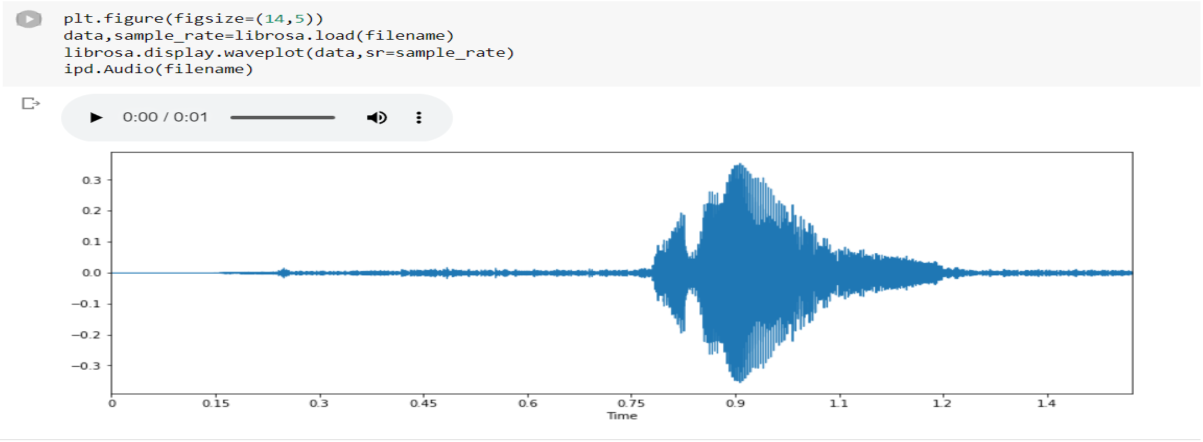
B. Class Distributions:

We selected 100 audios including 37 audios for person, 35 audios for location,14 audios for organization and 14 audios for date.

80 audios were used for training and 20 audios were used for testing.

C. Audio Data overview and Analysis

All audio samples are in .wav format and are sampled at discrete periods of time at the standard sampling rate (44.1kHz meaning 44,100 samples per second).



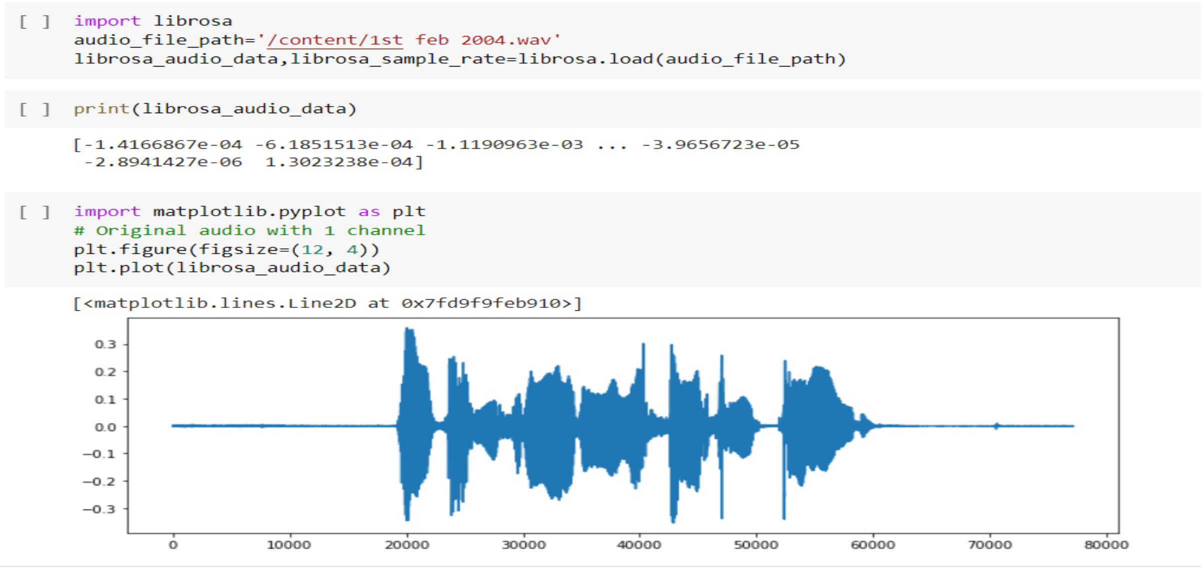
```
plt.figure(figsize=(14,5))
data,sample_rate=librosa.load(filename)
librosa.display.waveplot(data,sr=sample_rate)
ipd.Audio(filename)
```

Figure 1: Audio of Class Person

The bit intensity determines how designated the pattern will be (generally 16-bit samples can variety to approximately sixty five thousand amplitude values) and each pattern is the amplitude of the wave at a specific example of time. Therefore, the facts we are able to be the usage of for each audio pattern is essentially a unidimensional vector of amplitude values.

D. Visual Inspection of audio samples

We tried loading a sample from each class and visually inspect the data for any similarities or patterns . We use librosa to load the sound files in an array and then use matplotlib.pyplot and librosa.display to visualise the audio wave.



```
[ ] import librosa
    audio_file_path='/content/1st feb 2004.wav'
    librosa_audio_data,librosa_sample_rate=librosa.load(audio_file_path)
```

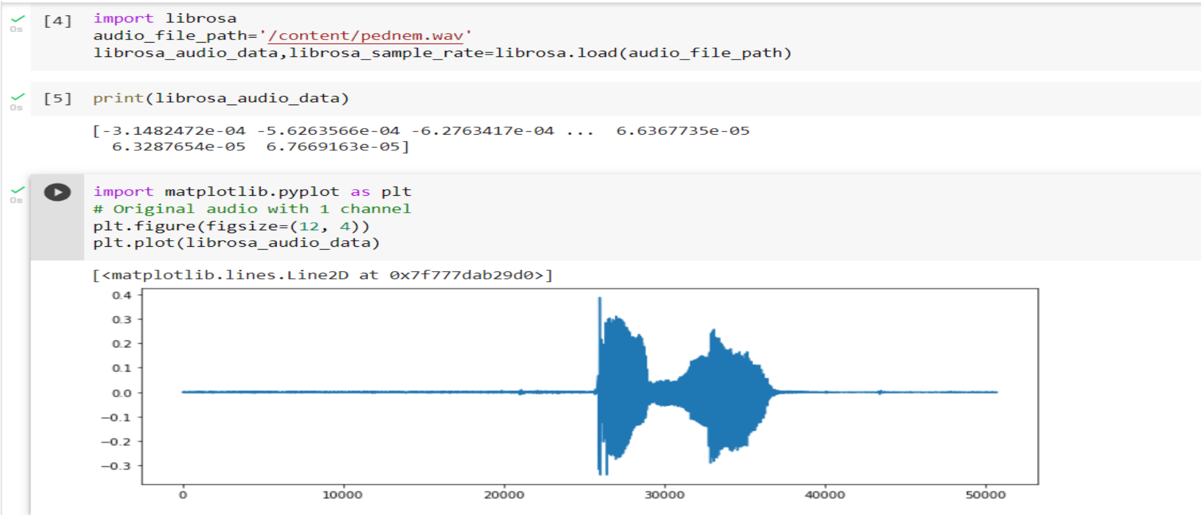
```
[ ] print(librosa_audio_data)

    [-1.4166867e-04 -6.1851513e-04 -1.1190963e-03 ... -3.9656723e-05
     -2.8941427e-06  1.3023238e-04]
```

```
[ ] import matplotlib.pyplot as plt
    # Original audio with 1 channel
    plt.figure(figsize=(12, 4))
    plt.plot(librosa_audio_data)

    [<matplotlib.lines.Line2D at 0x7fd9f9feb910>]
```

Figure 2: Audio of Class Date



```
[4] import librosa
    audio_file_path='/content/pednem.wav'
    librosa_audio_data,librosa_sample_rate=librosa.load(audio_file_path)

[5] print(librosa_audio_data)

    [-3.1482472e-04 -5.6263566e-04 -6.2763417e-04 ...  6.6367735e-05
      6.3287654e-05  6.7669163e-05]

    import matplotlib.pyplot as plt
    # Original audio with 1 channel
    plt.figure(figsize=(12, 4))
    plt.plot(librosa_audio_data)

    [<matplotlib.lines.Line2D at 0x7f777dab29d0>]
```

Figure 3: Audio of Class Location

One cannot simply visualize the differences between the audio classes just by visual inspection.

4. ACOUSTIC FEATURES FOR AUDIO CLASSIFICATION

4.1 Mel Frequency Cepstral Coefficients

MFCCs extraction is an essential goal of extracting the capabilities is to compress the audio signal to a vector this is consultant of the significant statistics it is attempting to characterize. In those works, acoustic capabilities particularly MFCC capabilities are extracted.

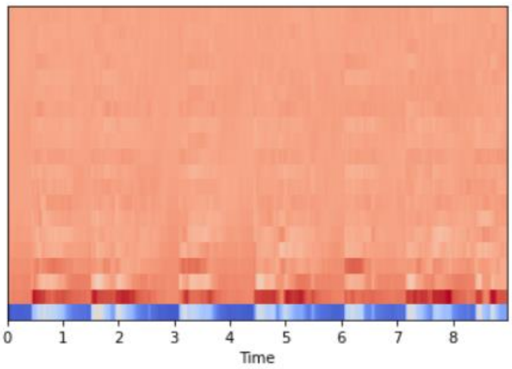


Figure 4: Mel- spectrogram

Mel Frequency Cepstral Coefficients (MFCCs) are short-time period spectral primarily based totally and dominant capabilities and are extensively used withinside the place of audio and speech processing. The mel frequency cepstrum has established to be rather powerful in spotting the shape of song alerts and in modeling the subjective pitch and frequency content material of audio alerts .

The MFCCs had been carried out in a variety of audio mining tasks, and feature proven desirable overall performance in comparison to different capabilities. MFCCs are computed with the aid of using diverse authors in exclusive methods. It computes the cepstral coefficients at the side of delta cepstral electricity and strength spectrum deviation which ends up in 26 dimensional capabilities.

The low order MFCCs consists of statistics of the slowly converting spectral envelope at the same time as the better order MFCCs explains the quick versions of the envelope .

MFCC 13 dimensional feature values will be calculated for the given wav file. The above process is continued for 100 wav files.

5. CLASSIFICATION MODELS

5.1 Cnn Model

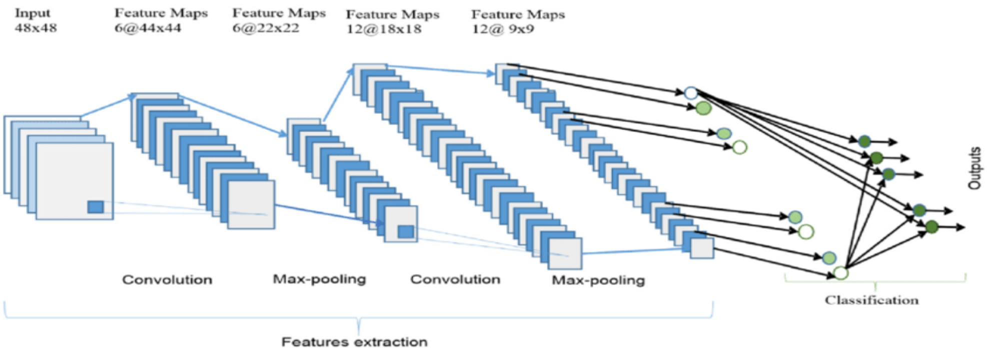


Figure 5: CNN Model

In a typical CNN, there is a series of various kind of layers which are combined in the overall architecture. The training of a CNN requires a different kind of decisions which need to be taken in terms of both architectural patterns such as number of convolution and pooling layers, input data format, filter dimension, etc, as well as hyperparameters such as learning rate, dropout probability, number of epochs, batch size, etc.

A convolutional neural community, or CNN, is a deep studying neural community sketched for processing based arrays of records inclusive of portrayals.

CNN are very quality at selecting up on layout withinside the enter photograph, inclusive of lines, gradients, circles, or maybe eyes and faces.

This feature that makes convolutional neural community so strong for laptop vision. CNN can run immediately on a underdone photograph and do now no longer want any preprocessing.

A convolutional neural community is a feed ahead neural community, seldom with as much as 20.

The strength of a convolutional neural community comes from a selected form of layer known as the convolutional layer. CNN consists of many convolutional layers assembled on pinnacle of every different, every one in a position of spotting extra state-of-the-art shapes.

With 3 or 4 convolutional layers it's miles possible to understand handwritten digits and with 25 layers it's miles viable to distinguish human faces.

The schedule for this sphere is to spark off machines to view the arena as human beings do, understand it in a like style or even use the information for a mess of obligationsinclusive of photograph and video recognition, photograph inspection and classification, media recreation, advice systems, herbal language processing, etc.

Convolutional Neural Network Design: The creation of a convolutional neural community is a multi-layered feed-ahead neural community, made with the aid of using assembling many unseen layers on pinnacle of every different in a selected order.

It is the sequential layout that supply permission to CNN to research hierarchical attributes.

In CNN, a number of them accompanied with the aid of using grouping layers and hidden layers are generally convolutional layers accompanied with the aid of using activation layers.

The pre-processing wanted in a ConvNet is kindred to that of the associated sample of neurons withinside the human mind and changed into influenced with the aid of using the employer of the Visual Cortex.

5.2 Support Vector Machine

A machine learning technique which is based on the principle of structure risk minimization is support vector machines. It has numerous applications in the area of pattern recognition [9]. SVM constructs linear model based upon support vectors in order to estimate decision function. If the training data are linearly separable, then SVM finds the optimal hyper plane that separates the data without error [10].

Figure 6 shows an example of a non-linear mapping of SVM to construct an optimal hyper plane of separation. SVM maps the input patterns through a non-linear mapping into higher dimension feature space. For linearly separable data, a linear SVM is used to classify the data sets [11]. The patterns lying on the margins which are maximized are the support vectors.

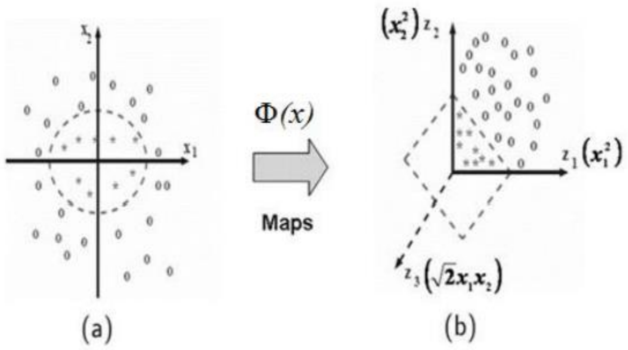


Figure -6: Example for SVM Kernel Function Φ(x) Maps 2- Dimensional Input Space to Higher 3-Dimensional Feature Space. (a) Nonlinear Problem. (b) Linear Problem.

The support vectors are the (transformed) training patterns and are equally close to hyperplane of separation. The support vectors are the training samples that define the optimal hyperplane and are the most difficult patterns to classify [12]. Informally speaking, they are the patterns most informative of the classification task. The kernel function generates the inner products to construct machines with different types of non-linear decision surfaces in the input space [13].

6. EXPERIMENTAL RESULTS

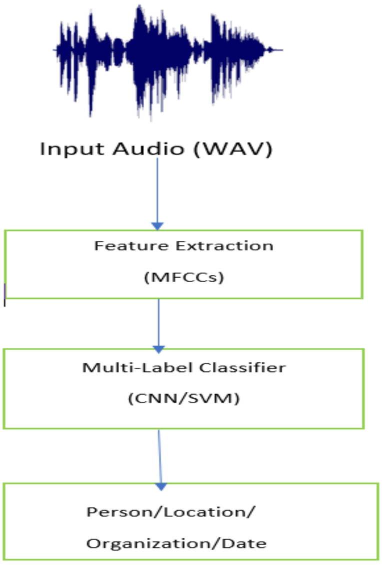


Figure 7: The proposed framework for Named entity Recognition For Konkani Speech

6.1 Dataset

The dataset(audios) used for this project is collected by recording the voices of different speakers. These datasets are then divided into four different categories i.e Person, Location, Date, and organization. The audios are stored in the WAV format as this format recreates the recording accurately without losing audio quality due to the format itself.

6.2 Feature Extraction

An input audio wav file is given as the input to the feature extraction techniques. In feature extraction technique MFCC features are extracted, and the MFCC features help to uniquely identify the audio signals.

MFCC 13 dimensional feature values will be calculated for the given wav file. The above process is continued for 100 wav files.

6.3 Classification

Once the feature extraction process is done, the named entities should be classified. We selected 100 audios for training data including 37 audios for person, 35 audios for location,14 audios for organization and 14 audios for date.

The Classification models used for the classification of named entities are CNN and SVM classifier.

The models are trained to classify the named entities into person, location, organization, date etc.

7.RESULT:

The evaluation metric for this project will be "Classification Accuracy" which is defined as the percentage of correct predictions.

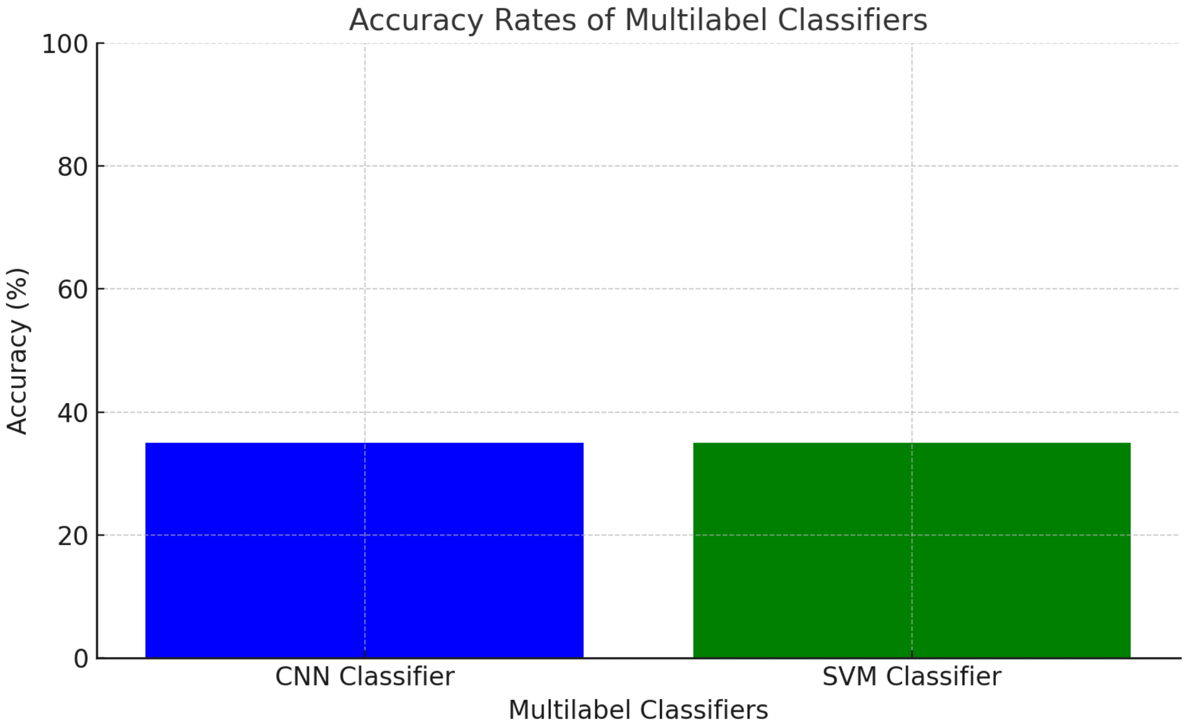
Accuracy = Correct Classifications/ Number of Classifications

Other metrics such as Precision, Recall were ruled out as they are more applicable to classification challenges.

The accuracy rate of the multilabel classifiers implemented are shown in the table below:

Sr. No.	Multi Label Classifiers	Accuracy
1.	CNN Classifier	35.0 %
2.	SVM Classifier	35.0 %

Table 1: Accuracy rate of the multilabel classifiers



Here is the graph representing the accuracy rates of the multilabel classifiers. It clearly shows that both the CNN and SVM classifiers achieved the same accuracy of 35%.

Both models achieved an accuracy of 35%. While this accuracy is relatively low, it is reasonable given that this is the first attempt at developing a Named Entity Recognition (NER) system for Konkani speech. Future efforts can focus on improving accuracy by increasing the dataset size, which will provide better training data for the models.

8. CONCLUSION:

In this paper, we have proposed Named Entity Recognition for Konkani Speech. MFCC is calculated as features to characterize audio content. These features extracted are then used for the classification of named entities into pre-defined categories such as a person, location, organization, and date. CNN is used for this multi-label classification process which gives an accuracy of 35 %. The classification was done again using the SVM model to check if the accuracy improves but even the SVM model gave the same accuracy i.e 35 %. The accuracy is quite less but since this was the first attempt to develop NER for Konkani Speech, this accuracy is acceptable. In the future, the performance and the accuracy of the model can be improved by rising the size of the dataset for training.

REFERENCES:

1) Hemant Yadav, Sreyan Ghosh, Yi Yu, Rajiv Ratn Shah "End-to-end Named Entity Recognition from English Speech" Interspeech-2020.
2) Ankita Pasad, Felix Wu, Suwon Shon, Karen Livescu, Kyu J. Han"On the Use of External Data for Spoken Named Entity Recognition" Interspeech-2020.
3) Pham Ngoc Phuong, Chung Tran Quang,Quang Minh Nguyen,Quoc Truong Do"Improving prosodic phrasing of Vietnamese text-to"2020.

4) Dejan Porjazovski, Juho Leinonen, Mikko Kurimo "Attention-Based End-to-End Named Entity Recognition from Speech"2020.

5) Athavale, V., Bhardwaj, S., Pamecha, M., Prabhu, A. Shrivastava, S., “Towards Deep Learning in Hindi NER: An approach to tackle the Labelled Data Scarcity,” In NLPAI, 2016, pp: 154-160.

6) K. S. Hasan, M. ur Rahman, and V. Ng, “Learning -Based Named Entity Recognition for Morphologically-Rich Resource-Scare Languages,” in Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, 2009, pp. 354–362.

7) V. R and S. L, “Domain focussed Named Entity Recognizer for Tamil using Conditional Random Fields,” in Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages, Hyderabad, India, 2008, pp. 59–66.

8) P.Srikanth and K. N. Murthy, “Named Entity Recognition for Telegu,” in Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages, Hyderabad, India, Jan 2008, pp. 41–50.

9) Chungsoo Lim Mokpo, Yeon-Woo Lee, and Joon-Hyuk Chang, “New Techniques for Improving the practicality of a SVM-Based Speech/Music Classifier,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1657-1660, 2012.

10) Hongchen Jiang, JunmeiBai, Shuwu Zhang, and Bo Xu, “SVM-Based Audio Scene Classification,” IEEE International Conference Natural Language Processing and Knowledge Engineering, Wuhan, China, pp. 131- 136, October 2005.

11) Lim and Chang, “Enhancing Support Vector MachineBased Speech/Music Classification using Conditional Maximum a Posteriori Criterion,” Signal Processing, IET, vol. 6, no. 4, pp. 335-340, 2012.
12) Md. Al Mehedi Hasan and Shamim Ahmad. predSuccSite: Lysine Succinylation Sites Prediction in Proteins by using Support Vector Machine and Resolving Data Imbalance Issue. International Journal of Computer Applications 182(15):8-13, September 2018.
13) Hend Ab. ELLaban, A A Ewees and Elsaeed E AbdElrazek. A Real-Time System for Facial Expression Recognition using Support Vector Machines and k-Nearest Neighbor Classifier. International Journal of Computer Applications 159(8):23-29, February 2017.