Denoising Word Embeddings Generated from Social Media Text

Guided by:

Dr. Premjith B
Asst. Professor
CEN , Amrita School of Engineering

Co-Guided by :

Dr. Sanjanasri.J.P
Asst. Professor
CEN , Amrita School of Engineering

Presented by:

P H V Pavan Kumar
CB.EN.P2CEN20020

AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

INTRODUCTION

- India is a multilingual country where we often spot conversations on social media platforms like YouTube, Facebook and, Twitter in code-mixed text.
- The computational study of people's opinions, attitudes and emotions toward an entity is called Sentiment Analysis.
- In this work, Sentiment Analysis was conducted on CodeMix Dravidian language texts taken from social media conversations.
- To classify the underlying sentiments of text as positive, negative, mixed feelings, Native and non-Native, initially, we performed sentiment analysis using three different Deep-learning based architectures.
- It's frequent to notice that social media conversations frequently have typos and non-standard spelling and grammar variations in their text.

INTRODUCTION

- Word embeddings generated from state-of-the-art (SOTA) pre-trained models were also inherent with noise.

- The source for noise in the word embeddings can be:
  - Due to the typos in corpora.
  - During the embeddings generation process:
    - Embedding initialization.
    - Min-batch ordering.

- Various signal denoising algorithm techniques are applied to denoise the embeddings.

- The efficacy of the denoised embeddings is evaluated extrinsically using ML classifiers.

Objective

1. To develop deep learning models for sentiment analysis in Dravidian languages.
2. To denoise the word embeddings generated from social media conversations using various signal denoising algorithms.
3. To study the efficacy of denoised embeddings through extrinsic evaluation by considering sentiment analysis in Dravidian languages as a use case.

Dataset Description

- Dataset used in conducting experiments is taken from the Competition **Dravidian-CodeMix-FIRE2021**.

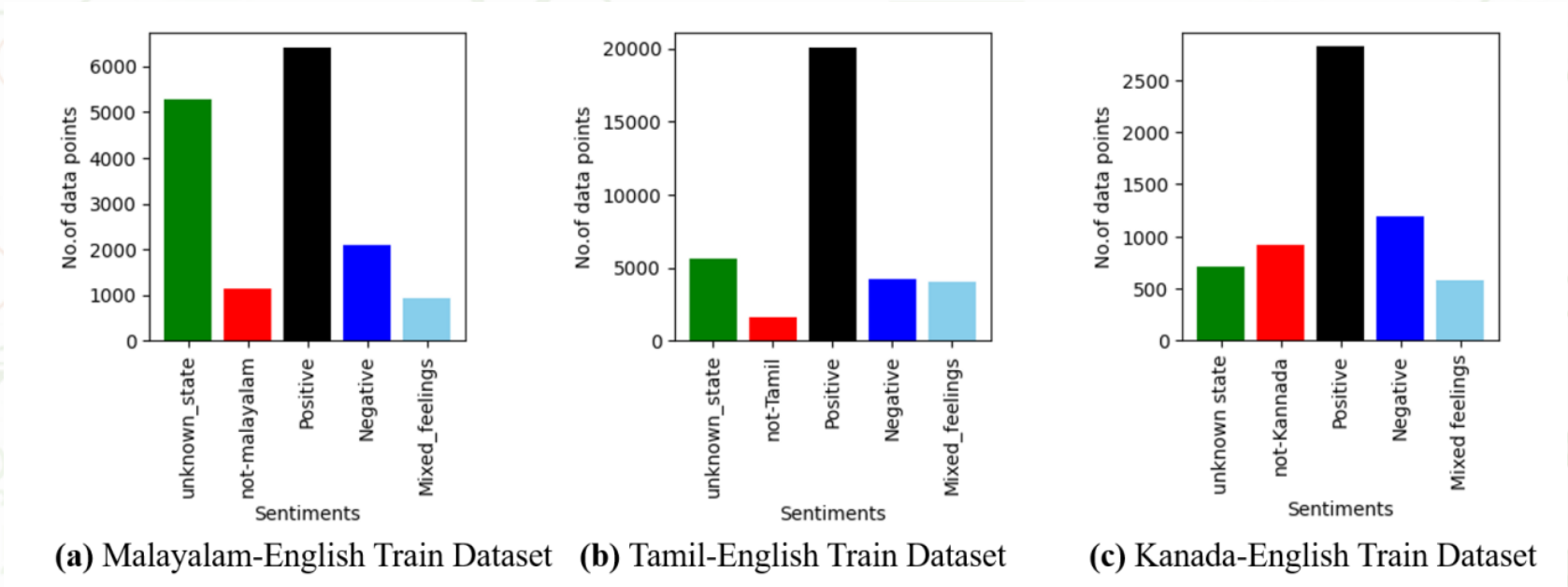
Unbalanced data (Phase 1)

Language	Class	Train	Validation	Test
Malayalam-English	unknown_state	15888	1766	1962
	Positive			
	Negative			
	Mixed_feelings			
	not-Malayalam			
Tamil-English	unknown_state	35656	3962	4402
	Positive			
	Negative			
	Mixed_feelings			
	not-Tamil			
Kanada-English	unknown state	6212	691	768
	Positive			
	Negative			
	Mixed feelings			
	not-Kannada			

Balanced data data (Phase 2)

Sentiments	Malayalam-English		Kanada-English		Tamil-English	
	Train	Test	Train	Test	Train	Test
unknown state	900	100	550	50	1600	**150**
Positive	900	100	550	50	1600	**150**
Negative	900	100	550	50	1600	**150**
Mixed feelings	900	100	550	50	1600	**150**
not-Dravidian	900	100	550	50	1600	**150**
Total	**4500**	**500**	**2750**	**250**	**8000**	**750**

Dataset Description (Phase 1)



(a) Malayalam-English Train Dataset **(b)** Tamil-English Train Dataset **(c)** Kanada-English Train Dataset

- Dataset is highly imbalanced, concept of class weights is applied to overcome this issue by computing the Individual class weights.

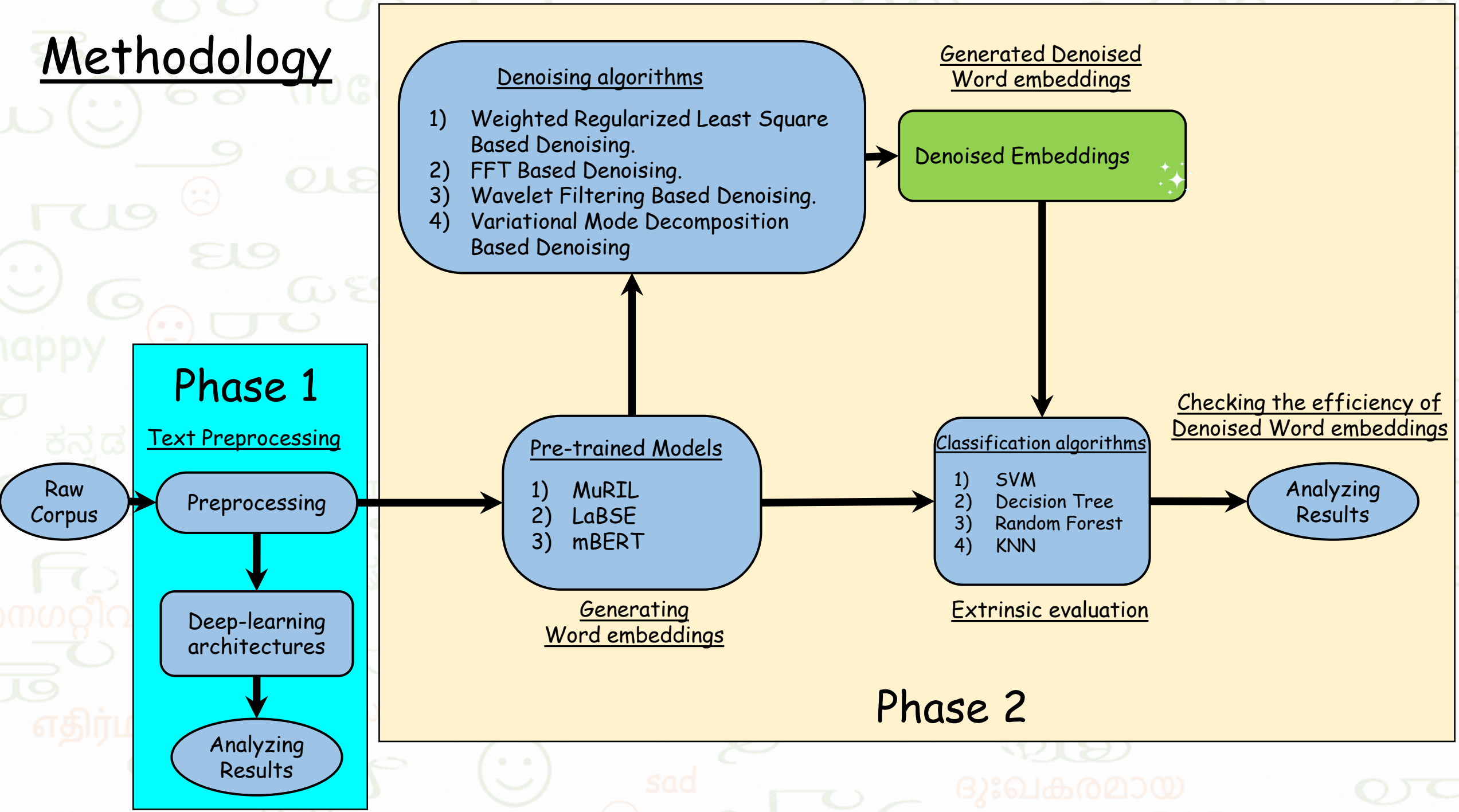
$$C_w = \frac{\sum_{c=1}^{n} N_c}{N_c}$$

$C_w \rightarrow$ *Class Weights*

$\sum_{c=1}^{n} N_c \rightarrow$ *Sum of all the sentences in the corpus.*

$N_c \rightarrow$ *Number of sentences in each class c.*

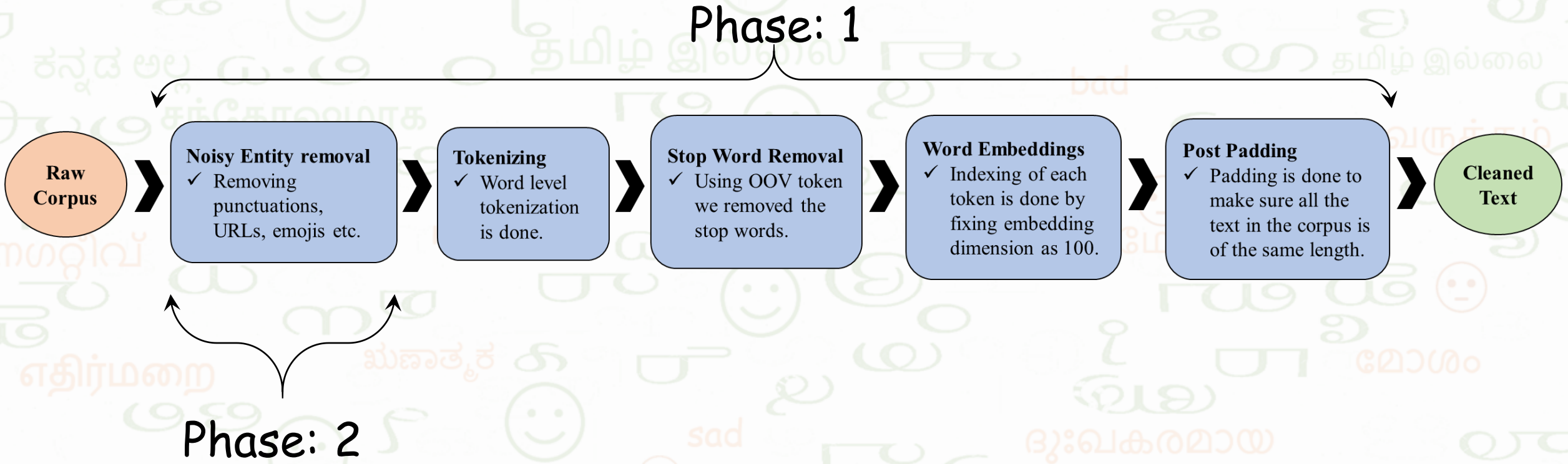
Methodology



Methodology (Phase 1)

Preprocessing:

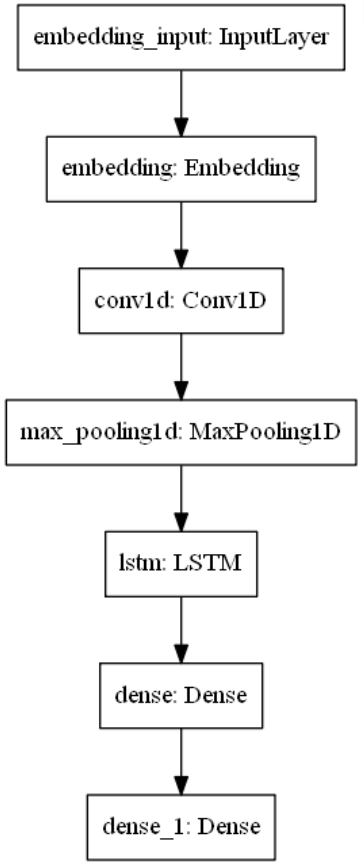
- Dataset contains lots of special Characters, emojis, URLs, and hashtags. These entities affect the performance of the Model accuracy. To remove all such entities from the corpus, we implemented the preprocessing stage.



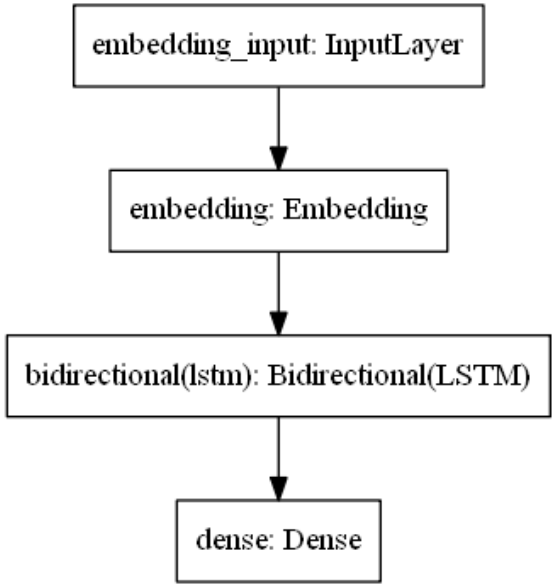
Methodology (Phase 1)

Description on Models:

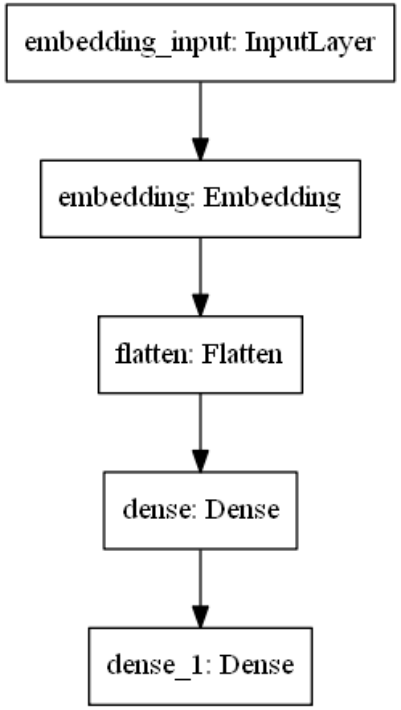
- By implementing three deep neural network architectures experiments had been conducted on the dataset.
- Each Model illustrated below follows a set of sequential steps before feeding into the network.



(a) Model-1



(b) Model-2



(c) Model-3

Methodology (Phase 1)

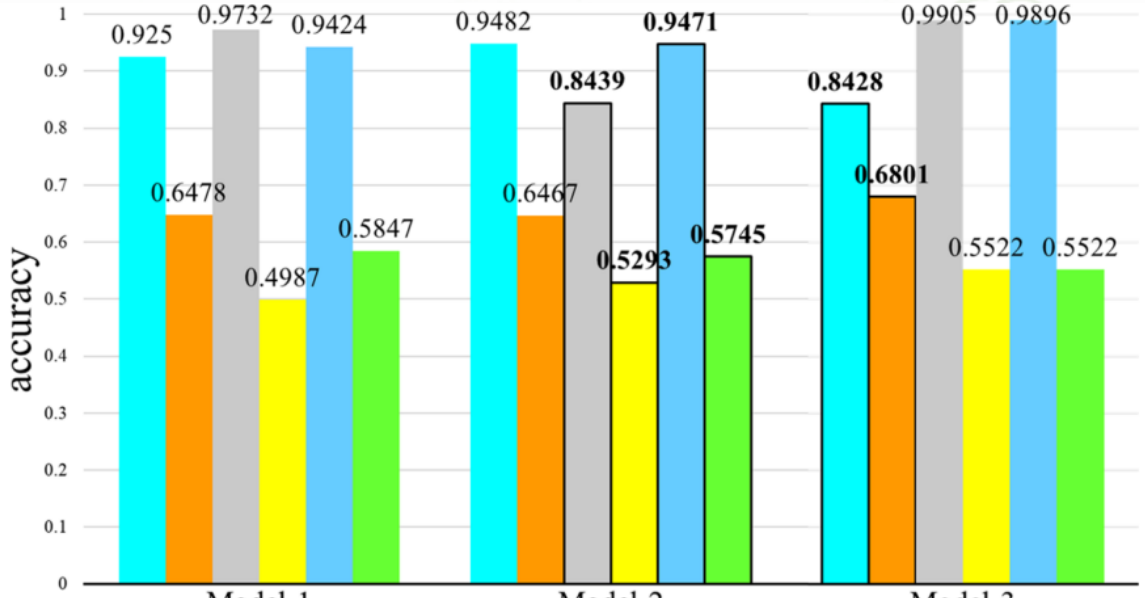
Hyperparameter tuning:

- Hyperparameter tuning was conducted based on improvements in Accuracy, Precision, Recall and, AUC values.
- Though the experiments are conducted on three models the hyperparameters of the best models are listed in the table.

Model 2	Embedding dimension	50, 100	100
	embeddings_initializer	uniform, orthogonal, constant	orthogonal
	embeddings_regularizer	L1, L2	L2
	Number of neurons in LSTM layer	16, 32, 64, 128, 256	32
	Activation Function at hidden layer	Sigmoid, RELU	RELU
	Activation Function at Output layer	Softmax	Softmax
	Optimizer	Adam	Adam
	Loss function	Sparse Categorical Crossentropy, Categorical Crossentropy	Categorical Crossentropy
	learning Rate	0.1, 0.01, 0.001	0.01
	Batch size	16, 32, 64, 80, 128, 132, 256	128
Model 3	Embedding dimension	50, 100	100
	Number of neurons in hidden layer	16, 32, 64, 128, 256	128
	Activation Function at hidden layer	Sigmoid, RELU	RELU
	Activation Function at Output layer	Softmax	Softmax
	Optimizer	Adam	Adam
	Loss function	Sparse Categorical Crossentropy, Categorical Crossentropy	Categorical Crossentropy
	learning Rate	0.1, 0.01, 0.001	0.01
	Batch size	16, 32, 64, 80, 128, 132, 256	64

Experiments and Results (Phase 1)

Testing Performance:



	Model-1	Model-2	Model-3
Malyalam-English Train accuracy	0.925	0.9482	**0.8428**
Malyalam-English Test accuracy	0.6478	0.6467	**0.6801**
Tamil-English Train accuracy	0.9732	**0.8439**	0.9905
Tamil- EnglishTest accuracy	0.4987	**0.5293**	0.5522
Kannada-English Train accuracy	0.9424	**0.9471**	0.9896
Kannada-English Test accuracy	0.5847	**0.5745**	0.5522

Language	Malayalam - English			Tamil - English			Kannada - English		
Model	**Precission**	**Recall**	**F1 Score**	**Precission**	**Recall**	**F1 Score**	**Precission**	**Recall**	**F1 Score**
Model-1	0.5854	0.6432	0.6077	0.4397	0.5072	0.4384	0.5007	0.5248	0.5085
Model-2	0.5797	0.6346	0.5995	**0.4232**	**0.5072**	**0.441**	**0.5062**	**0.5455**	**0.5193**
Model-3	**0.6303**	**0.6304**	**0.627**	0.43	0.4631	0.4408	0.4855	0.5126	0.4552

Conclusion (Phase 1)

- We did sentiment analysis for three Dravidian code-mixed languages, Malayalam, Tamil and, Kannada. We used three different deep learning Models:

1) Model-1 had a 1D-CNN layer, Maxpooling layer, LSTM, a fully connected dense layer.
2) Model-2 had one Bi-LSTM layer,
3) Model-3 had only one fully connected dense layer for conducting experiments.

After training three embedding Models on datasets several times, optimal hyperparameters were listed and the results obtained from Model-3 were much better when compared with Model-1 and Model-2 in Malayalam-English linguistics. Model-2 suits good for Kannada-English and Tamil-English linguistics

Denoising embeddings through extrinsic evaluation (Phase 2)

Pre-trained models used

- Being the dataset comprises three different CodeMix languages:

  a) Malayalam-English (Mal-Eng)
  b) Kannada-English (Kan-Eng)
  c) Tamil-English (Ta-Eng)

SOTA

- **Multilingual Representations for Indian Languages (MuRIL)** is a pre-trained on 17 Indian languages and their transliterated counterparts.
- **Language-agnostic BERT Sentence Embedding (LABSE)** pre-trained for sentence embedding for 109 languages.
- **Multilingual Bidirectional Encoder Representations from Transformers (mBERT)** pre-trained in 104 languages:

  1. BERT Cased: Model is case sensitive.
  2. BERT Uncased: Model is not case sensitive.

Weighted Regularized Least Square Based Denoising

Noisy Embedding

$$\min_x \left(\|y-x\|_2^2+\lambda\|Dx\|_2^2\right)$$

Formulating to Weighted regularized problem solving it.

$$x=(I+\lambda D^tD)^{-1}y$$

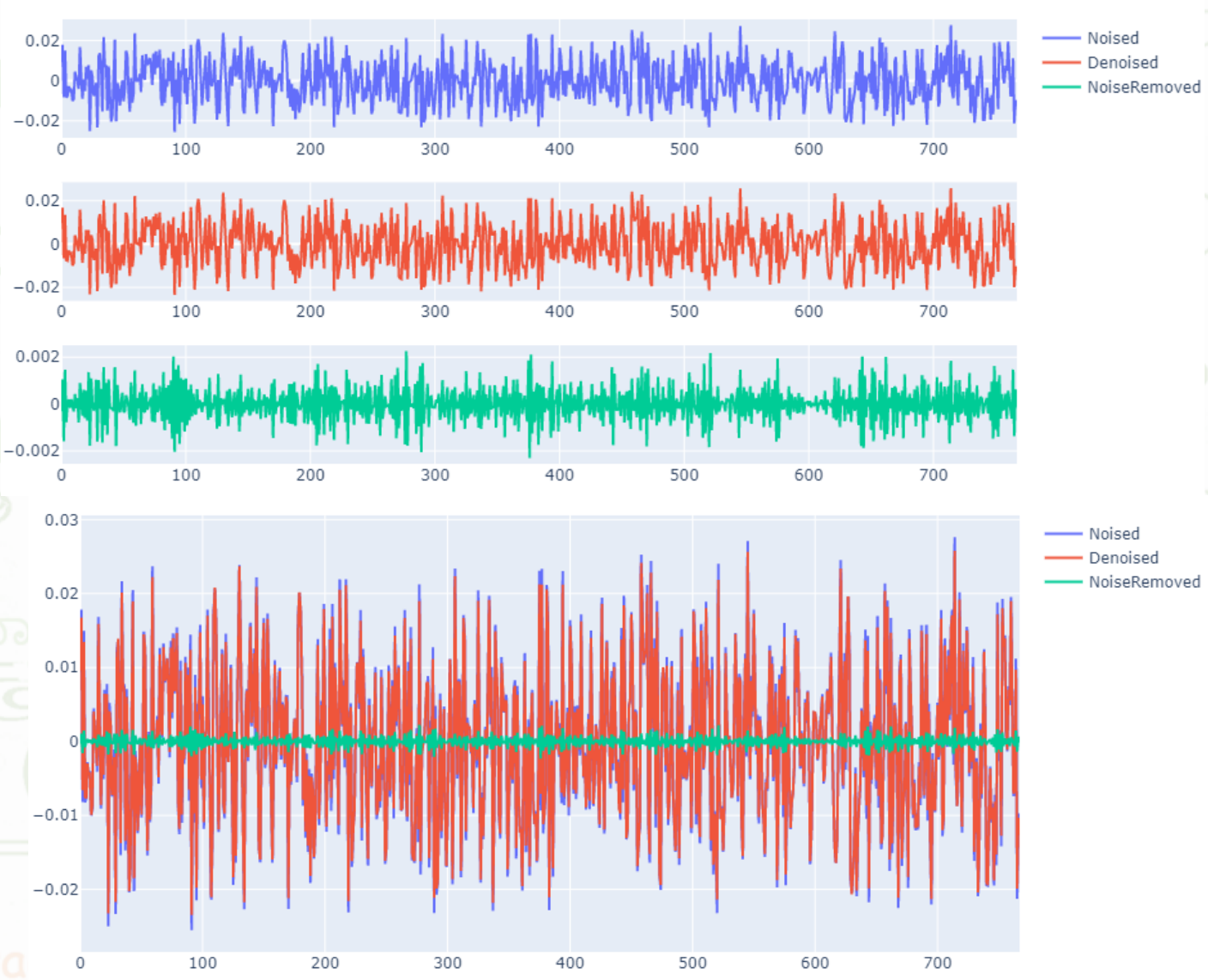
Noise free Embedding

$x \rightarrow denoised\ signal$

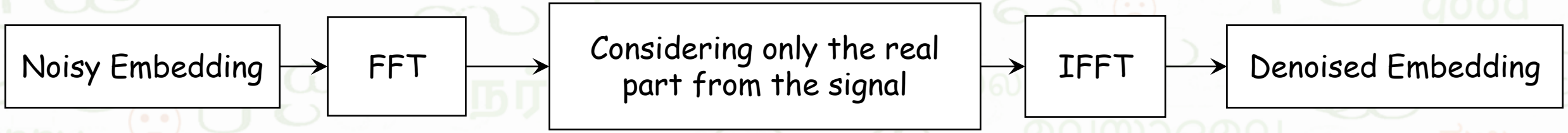
$y \rightarrow input\ signal$

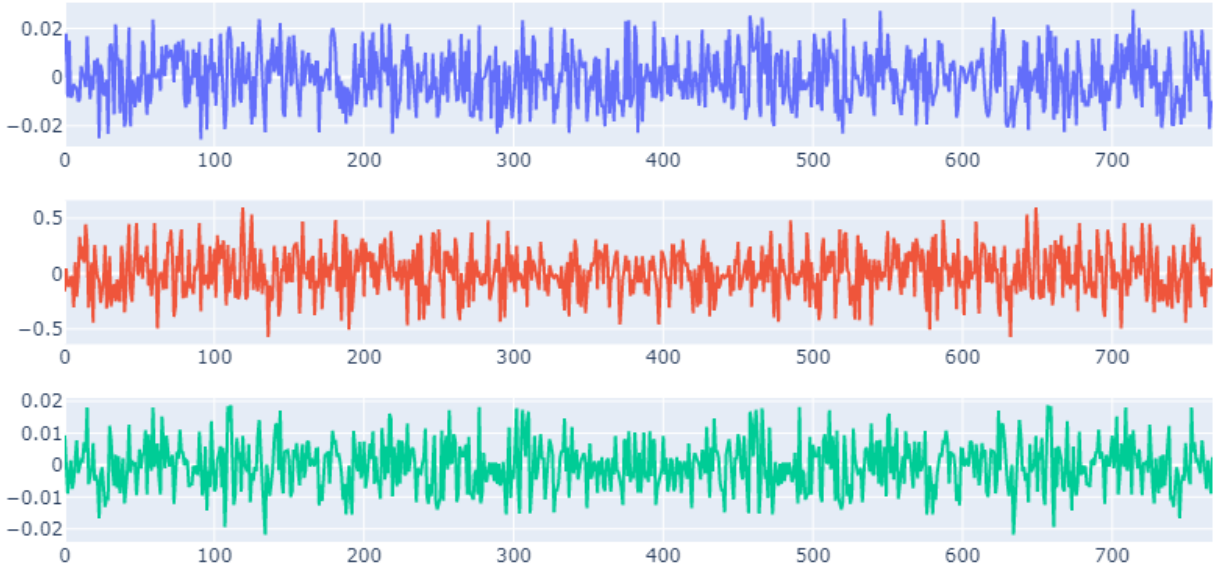
$D \rightarrow Order\ of Derivative$

$\boldsymbol{\lambda \rightarrow Hyper\ parameter}$

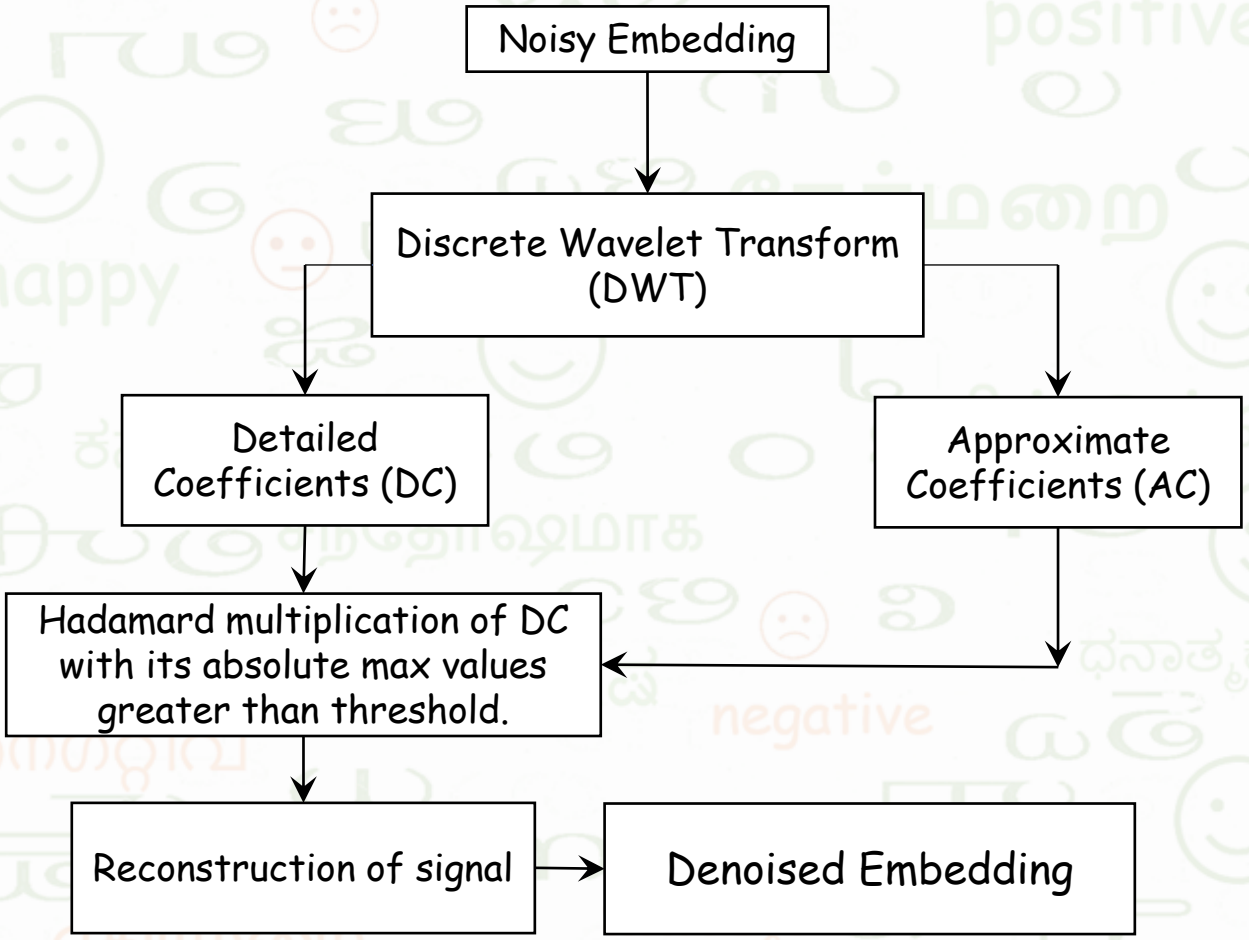


Fast Fourier Transform Denoising



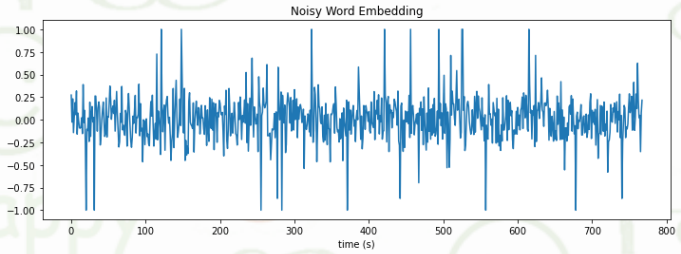


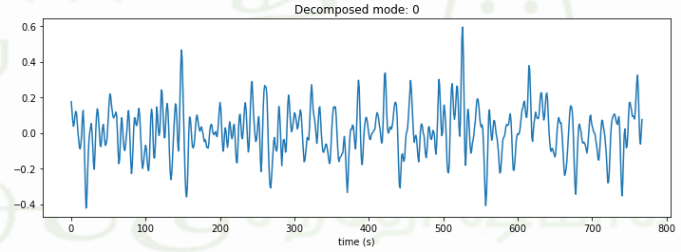
Wavelet-Based signal Denoising

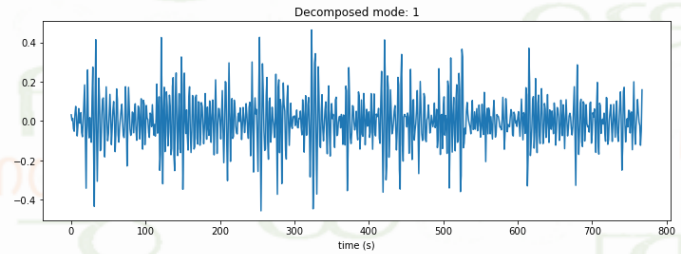


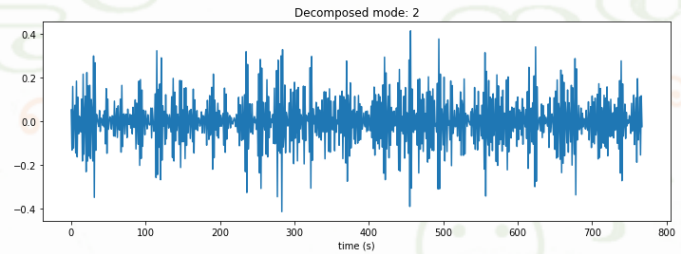
Variational Mode Decomposition (VMD) Denoising:

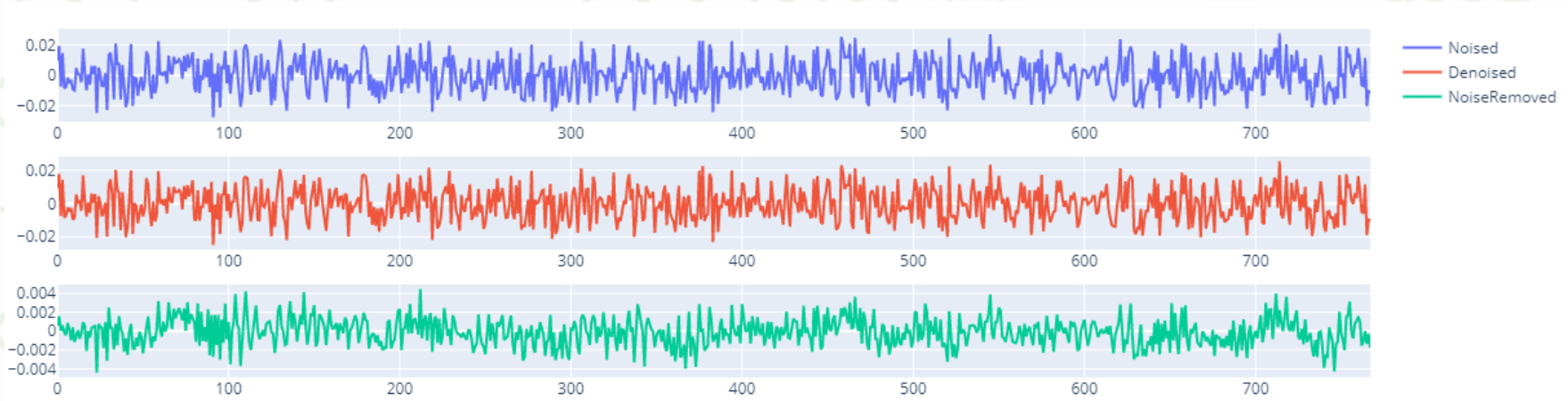
Noisy Embedding → VMD → Decomposing the signal → Reconstructing embedding → Denoised Embedding

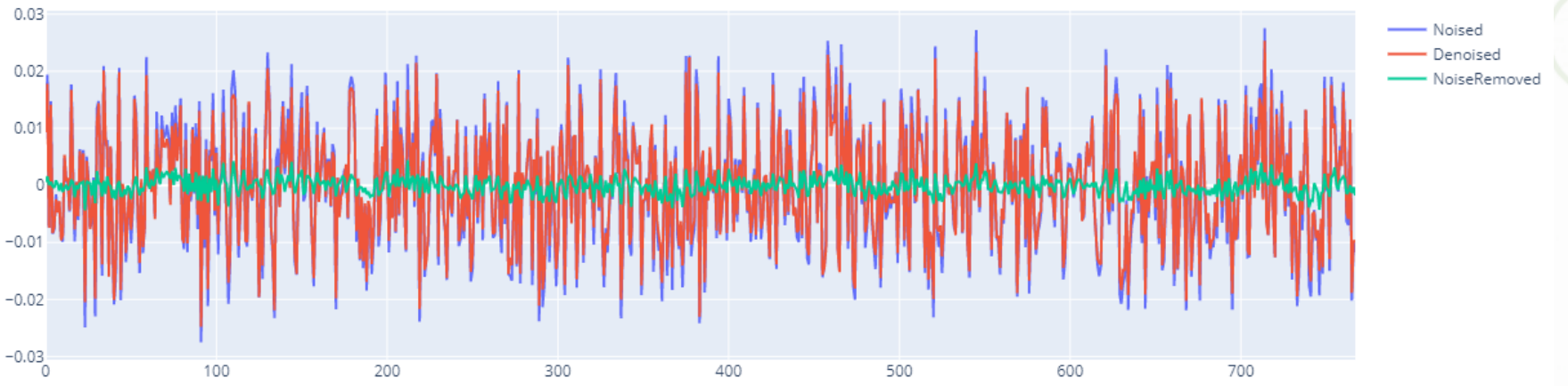












Hyper-parameters in VMD denoising:

Hyper-parameters	Description	Values
Alpha	The balancing parameter of the data-fidelity constraint.	100, 500, 1000, 2000, 3000, 5000
tau	Step on the recursion that updated the Lagrangian multipliers.	0, 0.1, 0.2, 0.3, 0.4, 0.5
Modes	The number of modes to be recovered.	1 to 50
DC	True if the first mode is put and kept at DC (0-freq).	0 and 1
init	The initialization of centre frequency. 0 = all omegas start at 0. 1 = all omegas start uniformly distributed. 2 = all omegas initialized randomly.	0, 1 and 2
Tol	Tolerance of convergence criterion.	1e-5, 1e-6
Mode selection	Modes used in reconstructing the signal.	All modes

Results:

Word level

Malayalam		BFR Denoising				WTR Denoising				FFT Denoisng				Wavlet Denoisng			
		Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
MuRIL	SVM	31.40%	43.05%	26.34%	31.40%	31.40%	43.05%	26.34%	31.40%	31.40%	43.04%	26.32%	31.40%	**32.40%**	**43.79%**	**27.13%**	**32.40%**
	DTree	34.00%	34.60%	34.25%	34.00%	**35.00%**	**34.87%**	**34.93%**	**35.00%**	**36.20%**	**35.67%**	**35.86%**	**36.20%**	29.20%	29.51%	29.34%	29.20%
	Rforest	37.80%	35.56%	36.25%	37.80%	**39.60%**	**37.64%**	**38.22%**	**39.60%**	**40.40%**	**38.81%**	**39.16%**	**40.40%**	36.40%	34.68%	35.24%	36.40%
	KNN	30.80%	30.22%	30.13%	30.80%	**31.20%**	**30.40%**	**30.42%**	**31.20%**	**32.60%**	**32.13%**	**31.81%**	**32.60%**	27.60%	26.84%	26.90%	27.60%
LABSE	SVM	51.20%	50.06%	50.38%	51.20%	51.20%	50.06%	50.39%	51.20%	**51.40%**	**50.19%**	**50.54%**	**51.40%**	51.20%	50.06%	50.38%	51.20%
	DTree	37.80%	37.71%	37.66%	37.80%	37.60%	37.47%	37.49%	37.60%	32.40%	32.34%	32.34%	32.40%	**39.40%**	**38.91%**	**38.70%**	**39.40%**
	Rforest	51.60%	50.17%	50.45%	51.60%	**50.80%**	**49.16%**	**49.66%**	**50.80%**	49.40%	47.57%	48.11%	49.40%	50.60%	49.44%	49.65%	50.60%
	KNN	51.40%	51.38%	50.54%	51.40%	**52.00%**	**51.83%**	**50.97%**	**52.00%**	49.80%	50.05%	49.21%	49.80%	51.00%	50.83%	50.14%	51.00%
BERT Cased	SVM	39.20%	46.25%	34.48%	39.20%	**39.40%**	**46.49%**	**34.71%**	**39.40%**	38.40%	39.75%	33.60%	38.40%	39.20%	46.25%	34.48%	39.20%
	DTree	34.80%	34.35%	34.54%	34.80%	34.60%	33.91%	34.20%	34.60%	31.00%	32.12%	31.43%	31.00%	32.20%	32.34%	32.26%	32.20%
	Rforest	45.00%	42.53%	43.10%	45.00%	45.00%	42.88%	43.49%	45.00%	**47.60%**	**45.18%**	**45.52%**	**47.60%**	43.80%	41.20%	41.78%	43.80%
	KNN	37.80%	38.06%	37.14%	37.80%	37.80%	38.21%	37.17%	37.80%	**38.20%**	**38.81%**	**37.61%**	**38.20%**	37.80%	38.06%	37.14%	37.80%
BERT Uncased	SVM	31.80%	44.70%	26.63%	31.80%	31.80%	44.70%	26.63%	31.80%	31.20%	43.99%	26.01%	31.20%	31.80%	44.70%	26.63%	31.80%
	DTree	32.80%	32.29%	32.50%	32.80%	32.20%	32.23%	32.19%	32.20%	**35.00%**	**34.48%**	**34.70%**	**35.00%**	**32.60%**	**32.30%**	**32.41%**	**32.60%**
	Rforest	42.80%	40.61%	41.27%	42.80%	**45.20%**	**42.83%**	**43.50%**	**45.20%**	**43.20%**	**41.41%**	**41.89%**	**43.20%**	**43.60%**	**41.45%**	**42.12%**	**43.60%**
	KNN	36.40%	36.25%	36.01%	36.40%	**37.00%**	**36.71%**	**36.56%**	**37.00%**	**37.20%**	**36.97%**	**36.91%**	**37.20%**	**36.60%**	**36.53%**	**36.25%**	**36.60%**

Note:

- If denoising performance is improving at W-level, then it has to be reflected at S-level as well, Further…

Results:

Sentence level

Malayalam		BFR Denoising				WTR Denoising				FFT Denoisng				Wavlet Denoisng			
		Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
MuRIL	SVM	33.60%	18.67%	21.47%	33.60%	**34.00%**	**19.59%**	**22.08%**	**34.00%**	**34.20%**	**19.64%**	**23.30%**	**34.20%**	**34.40%**	**20.31%**	**23.27%**	**34.40%**
	DTree	38.40%	37.14%	37.28%	38.40%	**42.00%**	**42.33%**	**41.96%**	**42.00%**	**40.00%**	**40.04%**	**39.95%**	**40.00%**	36.80%	37.79%	37.16%	36.80%
	Rforest	49.40%	47.71%	48.11%	49.40%	48.40%	46.90%	46.88%	48.40%	**51.00%**	**49.03%**	**49.58%**	**51.00%**	49.00%	48.07%	47.43%	49.00%
	KNN	48.60%	48.72%	47.42%	48.60%	47.20%	47.81%	46.14%	47.20%	**50.60%**	**51.38%**	**49.51%**	**50.60%**	48.40%	48.20%	47.40%	48.40%
LABSE	SVM	57.00%	56.91%	56.85%	57.00%	**57.60%**	**57.52%**	**57.46%**	**57.60%**	**57.60%**	**57.33%**	**57.41%**	**57.60%**	**57.20%**	**57.18%**	**57.08%**	**57.20%**
	DTree	33.20%	34.56%	33.68%	33.20%	**37.80%**	**37.63%**	**37.60%**	**37.80%**	**35.40%**	**36.09%**	**35.66%**	**35.40%**	**37.60%**	**40.09%**	**37.70%**	**37.60%**
	Rforest	49.40%	48.70%	48.73%	49.40%	**53.00%**	**52.21%**	**52.22%**	**53.00%**	**51.80%**	**50.92%**	**51.10%**	**51.80%**	**53.80%**	**53.38%**	**53.36%**	**53.80%**
	KNN	46.80%	51.00%	47.08%	46.80%	**47.40%**	**51.48%**	**47.78%**	**47.40%**	**46.80%**	**51.31%**	**47.27%**	**46.80%**	**47.00%**	**51.18%**	**47.30%**	**47.00%**
BERT Cased	SVM	46.40%	45.43%	45.00%	46.40%	46.20%	45.06%	44.58%	46.20%	**46.60%**	**46.19%**	**45.08%**	**46.60%**	**46.60%**	**45.69%**	**45.20%**	**46.60%**
	DTree	33.20%	33.30%	33.21%	33.20%	**38.20%**	**38.45%**	**38.26%**	**38.20%**	**35.60%**	**36.46%**	**35.93%**	**35.60%**	**34.40%**	**34.40%**	**34.36%**	**34.40%**
	Rforest	43.00%	41.63%	41.90%	43.00%	**45.20%**	**44.05%**	**44.43%**	**45.20%**	**44.00%**	**42.27%**	**42.64%**	**44.00%**	**43.80%**	**42.24%**	**42.71%**	**43.80%**
	KNN	34.40%	36.90%	34.58%	34.40%	34.40%	36.57%	34.38%	34.40%	**36.60%**	**38.62%**	**36.70%**	**36.60%**	34.60%	36.86%	34.64%	34.60%
BERT Uncased	SVM	38.40%	40.75%	33.63%	38.40%	38.40%	40.45%	33.60%	38.40%	37.40%	37.94%	32.48%	37.40%	38.00%	40.62%	33.03%	38.00%
	DTree	29.80%	29.71%	29.69%	29.80%	**32.20%**	**32.17%**	**32.13%**	**32.20%**	**32.00%**	**31.97%**	**31.96%**	**32.00%**	**30.20%**	**30.80%**	**30.40%**	**30.20%**
	Rforest	39.00%	37.36%	37.86%	39.00%	**42.00%**	**40.21%**	**40.60%**	**42.00%**	**43.20%**	**41.60%**	**42.00%**	**43.20%**	39.20%	36.58%	37.29%	39.20%
	KNN	33.40%	32.87%	32.30%	33.40%	32.80%	31.97%	31.56%	32.80%	**34.40%**	**33.75%**	**33.49%**	**34.40%**	32.60%	32.33%	31.66%	32.60%

Results:

Performance of Malayalam-English language at both sentence and word level.

Model	Classifier	WTR Denoising	FFT Denoising	Wavelet Denoising
MuRIL	SVM	✗	✗	✓
	Dtree	✓	✓	✗
	RForest	✗	✓	✗
	KNN	✗	✓	✗
LABSE	SVM	✗	✓	✗
	Dtree	✗	✗	✓
	RForest	✓	✗	✗
	KNN	✓	✗	✗
BERT Cased	SVM	✗	✗	✗
	Dtree	✗	✗	✗
	RForest	✗	✓	✗
	KNN	✗	✓	✗
BERT UnCased	SVM	✗	✗	✗
	Dtree	✗	✓	✓
	RForest	✓	✓	✗
	KNN	✗	✓	✗

✓ : Denoising happens at both word and sentence levels.
✗ : Denoising does not occur at both word and sentence levels together.

Results:

Word level

Kannada		BFR Denoising				WTR Denoising				FFT Denoisng				Wavlet Denoisng			
		Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
MuRIL	SVM	41.60%	44.38%	39.24%	41.60%	**42.00%**	**44.93%**	**39.76%**	**42.00%**	**43.60%**	**45.88%**	**41.51%**	**43.60%**	41.20%	41.34%	39.80%	41.20%
	DTree	41.60%	41.26%	36.30%	41.60%	40.80%	41.25%	38.14%	40.80%	40.80%	45.09%	39.76%	40.80%	41.20%	39.02%	35.76%	41.20%
	Rforest	44.80%	44.09%	44.15%	44.80%	**45.60%**	**44.60%**	**44.62%**	**45.60%**	**46.80%**	**45.63%**	**45.61%**	**46.80%**	**47.20%**	**46.09%**	**45.93%**	**47.20%**
	KNN	44.40%	45.62%	44.44%	44.40%	44.00%	45.35%	44.10%	44.00%	44.00%	44.60%	43.91%	44.00%	42.80%	43.67%	43.03%	42.80%
LABSE	SVM	48.00%	49.97%	47.16%	48.00%	47.60%	49.72%	46.82%	47.60%	46.40%	48.72%	45.58%	46.40%	48.00%	49.97%	47.16%	48.00%
	DTree	39.20%	39.47%	39.21%	39.20%	36.00%	36.65%	36.19%	36.00%	36.00%	36.47%	36.17%	36.00%	34.80%	34.55%	34.52%	34.80%
	Rforest	51.20%	51.36%	50.96%	51.20%	**52.40%**	**53.27%**	**52.17%**	**52.40%**	**51.60%**	**51.39%**	**51.10%**	**51.60%**	50.00%	51.01%	49.72%	50.00%
	KNN	48.00%	48.45%	47.28%	48.00%	47.20%	47.68%	46.57%	47.20%	45.60%	45.51%	45.10%	45.60%	46.40%	46.52%	45.77%	46.40%
BERT Cased	SVM	35.60%	34.15%	32.15%	35.60%	**48.40%**	**48.03%**	**47.78%**	**48.40%**	**37.60%**	**37.07%**	**34.45%**	**37.60%**	34.80%	33.41%	31.40%	34.80%
	DTree	35.20%	36.67%	35.71%	35.20%	**47.20%**	**49.95%**	**46.26%**	**47.20%**	**40.00%**	**40.16%**	**39.86%**	**40.00%**	**40.80%**	**44.19%**	**39.00%**	**40.80%**
	Rforest	46.00%	45.99%	45.63%	46.00%	**46.80%**	**47.12%**	**45.26%**	**46.80%**	**46.80%**	**46.26%**	**46.27%**	**46.80%**	**49.20%**	**49.16%**	**48.08%**	**49.20%**
	KNN	37.20%	39.79%	37.90%	37.20%	36.00%	38.52%	36.67%	36.00%	**37.60%**	**40.41%**	**38.04%**	**37.60%**	36.00%	38.42%	36.47%	36.00%
BERT Uncased	SVM	50.40%	51.56%	50.32%	50.40%	**50.80%**	**51.94%**	**50.85%**	**50.80%**	**52.80%**	**53.15%**	**52.74%**	**52.80%**	**52.80%**	**52.37%**	**51.72%**	**52.80%**
	DTree	42.00%	42.52%	42.02%	42.00%	42.80%	42.87%	41.10%	42.80%	**42.40%**	**41.93%**	**40.64%**	**42.40%**	42.00%	45.31%	41.27%	42.00%
	Rforest	48.00%	47.95%	47.62%	48.00%	**48.80%**	**47.99%**	**47.98%**	**48.80%**	47.60%	48.02%	47.22%	47.60%	46.40%	46.36%	45.78%	46.40%
	KNN	37.20%	38.53%	37.38%	37.20%	36.80%	38.30%	37.07%	36.80%	36.80%	37.54%	36.90%	36.80%	**40.40%**	**41.13%**	**40.10%**	**40.40%**

Results:

Sentence level

Kannada		BFR Denoising				WTR Denoising				FFT Denoisng				Wavlet Denoisng			
		Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
MuRIL	SVM	42.40%	43.36%	40.63%	42.40%	**51.60%**	**50.62%**	**50.84%**	**51.60%**	**48.80%**	**47.43%**	**47.60%**	**48.80%**	**46.40%**	**45.21%**	**45.61%**	**46.40%**
	DTree	38.40%	39.87%	38.37%	38.40%	**42.00%**	**41.75%**	**39.08%**	**42.00%**	**41.20%**	**41.37%**	**39.76%**	**41.20%**	**42.40%**	**40.77%**	**40.37%**	**42.40%**
	Rforest	48.80%	48.51%	48.25%	48.80%	**53.20%**	**52.95%**	**52.48%**	**53.20%**	**51.20%**	**51.17%**	**50.52%**	**51.20%**	**51.20%**	**51.37%**	**50.17%**	**51.20%**
	KNN	51.20%	52.78%	50.50%	51.20%	49.20%	50.72%	48.47%	49.20%	50.00%	51.77%	49.20%	50.00%	45.60%	46.27%	44.45%	45.60%
LABSE	SVM	58.00%	58.54%	57.88%	58.00%	57.60%	58.10%	57.47%	57.60%	53.60%	53.55%	53.31%	53.60%	56.40%	56.92%	56.25%	56.40%
	DTree	41.20%	42.89%	41.53%	41.20%	41.20%	41.02%	40.85%	41.20%	**42.80%**	**46.13%**	**42.28%**	**42.80%**	35.60%	36.44%	35.74%	35.60%
	Rforest	52.00%	52.18%	51.44%	52.00%	**54.40%**	**54.60%**	**54.09%**	**54.40%**	**55.60%**	**56.39%**	**55.27%**	**55.60%**	46.40%	47.01%	46.14%	46.40%
	KNN	47.20%	49.42%	46.65%	47.20%	46.40%	49.40%	45.58%	46.40%	46.00%	47.92%	45.24%	46.00%	46.80%	50.54%	45.73%	46.80%
BERT Cased	SVM	47.60%	46.67%	45.85%	47.60%	46.40%	45.26%	44.77%	46.40%	46.40%	45.40%	44.75%	46.40%	**49.60%**	**48.91%**	**47.82%**	**49.60%**
	DTree	31.20%	31.15%	31.13%	31.20%	**32.80%**	**32.90%**	**32.83%**	**32.80%**	**33.60%**	**33.31%**	**33.03%**	**33.60%**	30.00%	30.31%	30.07%	30.00%
	Rforest	44.00%	43.58%	42.90%	44.00%	42.80%	42.12%	41.56%	42.80%	**46.00%**	**45.65%**	**45.03%**	**46.00%**	**46.80%**	**46.81%**	**46.41%**	**46.80%**
	KNN	40.80%	44.14%	39.48%	40.80%	40.40%	40.77%	38.96%	40.40%	39.20%	43.26%	37.82%	39.20%	41.20%	43.55%	40.19%	41.20%
BERT Uncased	SVM	39.20%	32.29%	34.66%	39.20%	39.20%	32.47%	34.74%	39.20%	39.20%	32.28%	34.63%	39.20%	38.00%	34.01%	34.13%	38.00%
	DTree	30.40%	30.33%	30.22%	30.40%	**35.60%**	**35.00%**	**35.21%**	**35.60%**	27.60%	28.19%	27.54%	27.60%	**31.60%**	**32.07%**	**31.68%**	**31.60%**
	Rforest	43.20%	42.59%	42.65%	43.20%	**45.20%**	**45.32%**	**45.02%**	**45.20%**	**46.40%**	**45.74%**	**45.67%**	**46.40%**	**43.60%**	**44.33%**	**43.46%**	**43.60%**
	KNN	38.00%	38.77%	37.53%	38.00%	36.80%	37.89%	36.53%	36.80%	**38.80%**	**39.22%**	**38.26%**	**38.80%**	36.40%	34.11%	34.16%	36.40%

Results:

Performance of Kannada-English language at both sentence and word level.

Model	Classifier	WTR Denoising	FFT Denoising	Wavelet Denoising
MuRIL	SVM	✓	✓	✗
	Dtree	✗	✗	✗
	RForest	✓	✓	✓
	KNN	✗	✗	✗
LABSE	SVM	✗	✗	✗
	Dtree	✗	✗	✗
	RForest	✓	✓	✗
	KNN	✗	✗	✗
BERT Cased	SVM	✗	✗	✗
	Dtree	✓	✓	✗
	RForest	✗	✓	✓
	KNN	✗	✗	✗
BERT UnCased	SVM	✗	✗	✗
	Dtree	✗	✗	✗
	RForest	✓	✗	✗
	KNN	✗	✗	✗

✓ : Denoising happens at both word and sentence levels.
✗ : Denoising does not occur at both word and sentence levels together.

Results:

Word level

Tamil		BFR Denoising				WTR Denoising				FFT Denoisng				Wavlet Denoisng			
		Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
MuRIL	SVM	25.60%	21.60%	21.99%	25.60%	**25.73%**	**21.71%**	**22.11%**	**25.73%**	25.60%	21.60%	21.99%	25.60%	**25.87%**	**21.98%**	**22.44%**	**25.87%**
	DTree	29.47%	30.04%	29.69%	29.47%	29.33%	28.73%	28.97%	29.33%	**26.67%**	**27.12%**	**26.85%**	**26.67%**	25.20%	25.44%	25.28%	25.20%
	Rforest	31.73%	30.41%	30.74%	31.73%	**32.53%**	**30.98%**	**31.46%**	**32.53%**	**32.67%**	**31.78%**	**31.94%**	**32.67%**	31.60%	30.75%	30.99%	31.60%
	KNN	24.80%	24.73%	24.42%	24.80%	24.53%	24.41%	24.13%	24.53%	23.60%	23.51%	23.14%	23.60%	24.67%	24.94%	24.29%	24.67%
LABSE	SVM	45.73%	44.63%	44.85%	45.73%	**46.00%**	**44.99%**	**45.18%**	**46.00%**	**47.07%**	**46.37%**	**46.51%**	**47.07%**	45.07%	44.06%	44.26%	45.07%
	DTree	34.67%	36.64%	34.40%	34.67%	**35.33%**	**35.84%**	**35.47%**	**35.33%**	34.53%	39.28%	33.72%	34.53%	34.67%	36.64%	34.40%	34.67%
	Rforest	44.00%	43.41%	43.52%	44.00%	43.60%	42.87%	43.01%	43.60%	**44.80%**	**44.29%**	**44.47%**	**44.80%**	42.93%	42.24%	42.18%	42.93%
	KNN	40.67%	41.60%	40.95%	40.67%	40.00%	40.92%	40.28%	40.00%	39.73%	41.02%	40.00%	39.73%	40.67%	41.60%	40.95%	40.67%
BERT Cased	SVM	35.33%	40.61%	32.46%	35.33%	35.33%	40.56%	32.43%	35.33%	35.33%	42.15%	32.40%	35.33%	35.33%	40.61%	32.46%	35.33%
	DTree	31.73%	32.21%	31.90%	31.73%	27.20%	27.58%	27.34%	27.20%	27.07%	27.39%	27.16%	27.07%	**32.80%**	**32.78%**	**32.74%**	**32.80%**
	Rforest	39.33%	38.59%	38.70%	39.33%	39.33%	38.44%	38.63%	39.33%	**40.40%**	**39.46%**	**39.55%**	**40.40%**	**39.47%**	**38.77%**	**38.93%**	**39.47%**
	KNN	33.87%	36.51%	34.25%	33.87%	**34.40%**	**36.83%**	**34.73%**	**34.40%**	32.13%	34.36%	32.15%	32.13%	33.87%	36.51%	34.25%	33.87%
BERT Uncased	SVM	30.80%	25.18%	26.99%	30.80%	30.67%	25.10%	26.89%	30.67%	30.13%	24.82%	26.51%	30.13%	30.80%	25.18%	26.99%	30.80%
	DTree	27.20%	27.18%	27.16%	27.20%	**29.33%**	**29.45%**	**29.37%**	**29.33%**	27.20%	27.29%	27.24%	27.20%	24.67%	24.85%	24.68%	24.67%
	Rforest	35.20%	34.09%	34.22%	35.20%	35.07%	33.71%	33.99%	35.07%	**35.60%**	**33.99%**	**34.30%**	**35.60%**	**36.13%**	**35.00%**	**35.08%**	**36.13%**
	KNN	29.47%	31.06%	29.80%	29.47%	**29.87%**	**31.76%**	**30.23%**	**29.87%**	**30.67%**	**31.50%**	**30.65%**	**30.67%**	29.33%	30.92%	29.67%	29.33%

Results:

Sentence level

Tamil		BFR Denoising				WTR Denoising				FFT Denoisng				Wavlet Denoisng			
		Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall	Accuracy	Precision	F1	Recall
MuRIL	SVM	33.73%	47.14%	26.47%	33.73%	**45.46%**	**45.07%**	**45.13%**	**45.47%**	**44.80%**	**44.03%**	**44.31%**	**44.80%**	**44.80%**	**44.22%**	**44.37%**	**44.80%**
	DTree	34.53%	35.99%	34.92%	34.53%	31.20%	31.21%	31.19%	31.20%	27.20%	27.35%	27.25%	27.20%	32.00%	32.30%	31.55%	32.00%
	Rforest	43.33%	42.24%	42.47%	43.33%	42.40%	41.41%	41.52%	42.40%	40.53%	39.45%	39.67%	40.53%	40.80%	39.57%	39.20%	40.80%
	KNN	40.93%	41.46%	40.84%	40.93%	40.40%	41.09%	40.33%	40.40%	39.87%	40.40%	39.95%	39.87%	38.13%	38.48%	37.91%	38.13%
LABSE	SVM	47.20%	46.38%	46.51%	47.20%	**48.40%**	**47.80%**	**47.88%**	**48.40%**	**48.13%**	**47.55%**	**47.68%**	**48.13%**	47.20%	46.39%	46.51%	47.20%
	DTree	33.47%	34.15%	33.61%	33.47%	27.47%	27.54%	27.44%	27.47%	31.20%	31.29%	31.17%	31.20%	28.13%	28.03%	28.05%	28.13%
	Rforest	40.93%	40.56%	40.51%	40.93%	39.20%	38.66%	38.79%	39.20%	40.27%	39.31%	39.51%	40.27%	40.40%	39.85%	39.98%	40.40%
	KNN	40.53%	42.94%	41.17%	40.53%	39.47%	41.83%	40.07%	39.47%	39.60%	41.58%	40.19%	39.60%	40.53%	42.85%	41.14%	40.53%
BERT Cased	SVM	40.93%	40.15%	39.40%	40.93%	**41.20%**	**40.33%**	**39.67%**	**41.20%**	**42.27%**	**41.72%**	**41.66%**	**42.27%**	**41.07%**	**40.31%**	**39.63%**	**41.07%**
	DTree	26.40%	26.64%	26.48%	26.40%	**28.00%**	**28.28%**	**28.09%**	**28.00%**	**27.33%**	**27.53%**	**27.31%**	**27.33%**	**28.27%**	**28.02%**	**28.12%**	**28.27%**
	Rforest	38.00%	36.81%	36.59%	38.00%	**38.53%**	**37.05%**	**37.12%**	**38.53%**	36.40%	35.65%	35.35%	36.40%	38.00%	36.99%	37.04%	38.00%
	KNN	31.47%	33.17%	31.57%	31.47%	**31.87%**	**34.02%**	**32.08%**	**31.87%**	31.47%	33.97%	31.73%	31.47%	**31.87%**	**33.81%**	**32.03%**	**31.87%**
BERT Uncased	SVM	33.73%	30.36%	29.58%	33.73%	33.20%	29.88%	29.19%	33.20%	**37.60%**	**36.03%**	**36.04%**	**37.60%**	33.20%	30.36%	29.16%	33.20%
	DTree	27.47%	27.55%	27.51%	27.47%	**28.00%**	**28.30%**	**28.11%**	**28.00%**	**28.00%**	**27.48%**	**27.56%**	**28.00%**	27.33%	27.50%	27.40%	27.33%
	Rforest	36.93%	35.90%	36.01%	36.93%	**39.20%**	**38.30%**	**38.44%**	**39.20%**	**38.80%**	**38.25%**	**38.20%**	**38.80%**	35.07%	34.30%	34.51%	35.07%
	KNN	32.40%	33.52%	32.56%	32.40%	32.40%	33.30%	32.47%	32.40%	31.73%	33.09%	31.85%	31.73%	32.27%	33.31%	32.40%	32.27%

Results:

Performance of Tamil-English language at both sentence and word level.

Model	Classifier	WTR Denoising	FFT Denoising	Wavelet Denoising
MuRIL	SVM	✓	✗	✓
	Dtree	✗	✗	✗
	RForest	✗	✗	✗
	KNN	✗	✗	✗
LABSE	SVM	✓	✓	✗
	Dtree	✗	✗	✗
	RForest	✗	✗	✗
	KNN	✗	✗	✗
BERT Cased	SVM	✗	✗	✗
	Dtree	✗	✗	✓
	RForest	✗	✗	✗
	KNN	✓	✗	✗
BERT UnCased	SVM	✗	✗	✗
	Dtree	✓	✗	✗
	RForest	✗	✓	✗
	KNN	✗	✗	✗

✓ : Denoising happens at both word and sentence levels.

✗ : Denoising does not occur at both word and sentence levels together.

Results:

Malayalam embeddings from BERT Cased (Extrinsic evaluation)

BERT Cased					
Malayalam		**Accuracy**	**Precision**	**F1**	**Recall**
Bfr Denoising	**SVM**	46.40%	45.43%	45.00%	46.40%
	DTree	33.20%	33.30%	33.21%	33.20%
	Rforest	43.00%	41.63%	41.90%	43.00%
	KNN	34.40%	36.90%	34.58%	34.40%
WTR Denoising	**SVM**	46.20%	45.06%	44.58%	46.20%
	DTree	**38.20%**	**38.45%**	**38.26%**	**38.20%**
	Rforest	*45.20%*	*44.05%*	*44.43%*	*45.20%*
	KNN	34.40%	36.57%	34.38%	34.40%
FFT Denoising	**SVM**	46.60%	46.19%	45.08%	46.60%
	DTree	35.60%	36.46%	35.93%	35.60%
	Rforest	44.00%	42.27%	42.64%	44.00%
	KNN	**36.60%**	**38.62%**	**36.70%**	**36.60%**
Wavelet Denoising	**SVM**	46.60%	45.69%	45.20%	46.60%
	DTree	34.40%	34.40%	34.36%	34.40%
	Rforest	43.80%	42.24%	42.71%	43.80%
	KNN	34.60%	36.86%	34.64%	34.60%
VMD Denoising	**SVM**	**47.80%**	**47.54%**	**46.57%**	**47.80%**
	DTree	*36.20%*	*37.27%*	*36.65%*	*36.20%*
	Rforest	**45.40%**	**44.20%**	**44.46%**	**45.40%**
	KNN	*35.60%*	*38.35%*	*35.75%*	*35.60%*

Best Hyperparameter tunning

Malayalam		**alpha_tau_modes_DC_init_Tol_mode-selection**
VMD Denoising	**SVM**	**1000_0.1_8_0_1_1e-05all_modes**
	DTree	1000_0.1_10_0_1_1e-05all_modes
	Rforest	**1000_0.1_11_0_1_1e-05all_modes**
	KNN	1000_0.1_9_0_1_1e-05all_modes

- Among all classifiers, SVM with VMD denoised embeddings shows a **1.37% improvement** in the classification performance compared with all traditional denoising algorithms.
- The Rforest show **0.03% improvement**, which is not a good sign for denoising.
- DTree and KNN show no sign of denoising with VMD denoising.

Results:

Kannada embeddings from MuRIL (Extrinsic evaluation)

MuRIL					
Kannada		**Accuracy**	**Precision**	**F1**	**Recall**
Bfr Denoising	**SVM**	48.40%	47.42%	47.38%	48.40%
	DTree	38.40%	39.87%	38.37%	38.40%
	Rforest	48.80%	48.51%	48.25%	48.80%
	KNN	51.20%	52.78%	50.50%	51.20%
WTR Denoising	**SVM**	*51.60%*	*50.62%*	*50.84%*	*51.60%*
	DTree	*42.00%*	*41.75%*	*39.08%*	*42.00%*
	Rforest	**53.20%**	**52.95%**	**52.48%**	**53.20%**
	KNN	49.20%	50.72%	48.47%	49.20%
FFT Denoising	**SVM**	**48.80%**	**47.43%**	**47.60%**	**48.80%**
	DTree	*41.20%*	*41.37%*	*39.76%*	*41.20%*
	Rforest	*51.20%*	*51.17%*	*50.52%*	*51.20%*
	KNN	50.00%	51.77%	49.20%	50.00%
Wavelet Denoising	**SVM**	*46.40%*	*45.21%*	*45.61%*	*46.40%*
	DTree	*42.40%*	*40.77%*	*40.37%*	*42.40%*
	Rforest	*51.20%*	*51.37%*	*50.17%*	*51.20%*
	KNN	45.60%	46.27%	44.45%	45.60%
VMD Denoising	**SVM**	*43.60%*	*44.72%*	*41.86%*	*43.60%*
	DTree	**43.20%**	**44.08%**	**43.55%**	**43.20%**
	Rforest	*52.80%*	*53.30%*	*52.47%*	*52.80%*
	KNN	**52.00%**	**53.66%**	**51.49%**	**52.00%**

Best Hyperparameter tunning

Kannada		**alpha_tau_modes_DC_init_Tol_mode-selection**
VMD Denoising	**SVM**	3000_0.1_28_0_1_1e-05all_modes
	DTree	**500_0_7_1_1_1e-06all_modes**
	Rforest	500_0.5_7_0_1_1e-06all_modes
	KNN	**5000_0.1_27_0_1_1e-05all_modes**

- Among the VMD denoising embeddings, DTree shows a **3.18% improvement in the F1-score** compared with DTree on Wavelet Denoising.
- VMD denoised embeddings with KNN are also a case where we can see an **improvement in the F1-score as 1.26%** compared with noised embedding (Bfr Denoising).
- SVM and Rforest models show no improvement with VMD denoised signals compared to the traditional denoising algorithms.

Results:

Tamil embeddings from LABSE (Extrinsic evaluation)

LABSE					
Tamil		**Accuracy**	**Precision**	**F1**	**Recall**
Bfr Denoising	**SVM**	47.20%	46.38%	46.51%	47.20%
	DTree	33.47%	34.15%	33.61%	33.47%
	Rforest	40.93%	40.56%	40.51%	40.93%
	KNN	40.53%	42.94%	41.17%	40.53%
WTR Denoising	**SVM**	**48.40%**	**47.80%**	**47.88%**	**48.40%**
	DTree	27.47%	27.54%	27.44%	27.47%
	Rforest	39.20%	38.66%	38.79%	39.20%
	KNN	39.47%	41.83%	40.07%	39.47%
FFT Denoising	**SVM**	*48.13%*	*47.55%*	*47.68%*	*48.13%*
	DTree	31.20%	31.29%	31.17%	31.20%
	Rforest	40.27%	39.31%	39.51%	40.27%
	KNN	39.60%	41.58%	40.19%	39.60%
Wavelet Denoising	**SVM**	47.20%	46.38%	46.51%	47.20%
	DTree	28.13%	28.03%	28.05%	28.13%
	Rforest	40.40%	39.85%	39.98%	40.40%
	KNN	40.53%	42.85%	41.14%	40.53%
VMD Denoising	**SVM**	47.33%	46.56%	46.65%	47.33%
	DTree	**35.20%**	**34.72%**	**34.91%**	**35.20%**
	Rforest	**45.40%**	**43.72%**	**44.17%**	**45.40%**
	KNN	40.53%	42.91%	41.17%	40.53%

Best Hyperparameter tunning

Tamil		**alpha_tau_modes_DC_init_Tol_mode-selection**
VMD Denoising	**SVM**	3000_0.1_20_0_1_1e-05all_modes
	DTree	**500_0.1_26_0_1_1e-05all_modes**
	Rforest	**500_0.1_30_0_1_1e-05all_modes**
	KNN	1000_0.1_16_0_1_1e-05all_modes

- We can notice that Rforest shows a **3.66% improvement** in the classification performance compared with all the traditional denoising algorithms.
- For the case of the traditional denoising algorithm, there is no performance improvement with the DTree classifier model. But with VMD denoising, there is a **1.31% improvement** in classification performance.
- SVM and KNN models show there is no sign of improvement with VMD denoising.

Observation

- Based on the experiments conducted, By applying the traditional signal denoising algorithms to the Sentence as we as on Word level Embeddings, there is an improvement in the classification performance. This clearly states that denoising is happening, and the noise in the social media text is proven.
- Applying the VMD denoising algorithm on the embeddings and evaluating at extrinsic levels. We notice that improvement in classification performance is not at a higher level.

Conclusion

- Initially, Sentiment Analysis was conducted on the dataset using the three different deep learning Models and evaluated its performance on each language.
- Experimental results evident that there exists noise in word embeddings.
- The best denoising algorithm for the corresponding languages are:

Language	Denoising algorithm
Malayalam-English	Fast Fourier Transform (FFT)
Kannada-English	Weighted Regularized Least Square Based Denoising (WTR)
Tamil-English	Weighted Regularized Least Square Based Denoising (WTR)

Repo Link

Publications:

Paper accepted:

- ✓ Pavan Kumar P.H.V, Premjith B, Sanjanasri J.P, Soman K.P *Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages*. FIRE 2021 Forum for Information Retrieval Evaluation Virtual Event.
- ✓ **Status:** Accepted and yet to be indexed
- ✓ **Publication Status:** In press

Further Publication:

- Pavan Kumar P.H.V, Premjith B, Sanjanasri J.P, *Denoising Word Embeddings Generated from Social Media Text.*
- **Status:** Final draft yet to be finalized.
- **Conference:** yet to be finalized

References

[1] Nguyen KA, Walde SS, Vu NT. Neural-based noise filtering from word embeddings. arXiv preprint arXiv:1610.01874. 2016 Oct 6.

[2] Caciularu A, Dagan I, Goldberger J. Denoising Word Embeddings by Averaging in a Shared Space. arXiv preprint arXiv:2106.02954. 2021 Jun 5.

[3] Malykh V, Logacheva V, Khakhulin T. Robust word vectors: Context-informed embeddings for noisy texts. InProceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text 2018 Nov (pp. 54-63).

[4] Guo Y, Tang Y. Collaborative filtering algorithm based on denoising auto-encoder and item embedding. In2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) 2017 Mar 25 (pp. 1751-1755). IEEE.

[5] Nooralahzadeh F, Lønning JT, Øvrelid L. Reinforcement-based denoising of distantly supervised NER with partial annotation. Association for Computational Linguistics.

References

[6] Veena PV, Devi GR, Sowmya V, Soman K. Least square based image denoising using wavelet filters. Indian journal of science and technology. 2016 Aug 18;9(30):1-6.

[7] Srikanth M, Krishnan KG, Sowmya V, Soman KP. Image denoising based on weighted regularized least square method. In2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT) 2017 Apr 20 (pp. 1-5). IEEE.

[8] Naveed K, Akhtar MT, Siddiqui MF, ur Rehman N. A statistical approach to signal denoising based on data-driven multiscale representation. Digital Signal Processing. 2021 Jan 1;108:102896.

[9] Li F, Verma S, Deng P, Qi J, Marfurt KJ. Seismic denoising using thresholded adaptive signal decomposition. InSEG Technical Program Expanded Abstracts 2017 2017 Aug 17 (pp. 5095-5099). Society of Exploration Geophysicists.

[10] Motwani MC, Gadiya MC, Motwani RC, Harris FC. Survey of image denoising techniques. InProceedings of GSPX 2004 Sep 27 (Vol. 27, pp. 27-30). Proceedings of GSPX.


positive
good
நேர்மறை
நல்ல
ஒళ್ళೆಯದು
happy
മലയാളമല്ല
ದುಃಖ
தமிழ் இல்லை
தமிழ் இல்லை
bad
சந்தோஷமாக
கெட்ட
வருத்தம்
ಧನಾತ್ಮಕ
negative
எதிர்மறை
ಋಣಾತ್ಮಕ
மோசம்
sad
ದುಃಖಕರಮಾಯ

Thank you