International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-14 Issue-6, May 2025

The Natural Language Processing Axioms in Classical Tamil for Zonal Dialects using Machine Learning

Perumal Sivarman, Prabaharan G, Senthil Kumar R

Abstract: ***Studying Natural Language Processing (NLP) for Classical Tamil and its Zonal Dialects using Machine Learning (ML) involves unique challenges and opportunities. Classical Tamil, one of the oldest languages with a rich literary heritage, differs significantly in syntax, semantics, and phonetics from its modern dialects. Addressing these differences requires incorporating linguistic axioms and cultural nuances into NLP systems. This deals with Tamil letters, Challenges, Future directions, and Lexical Differences. It also includes parsers and tokenisation. Lists the differences between Morphemes, Bounded Morphemes in terms of Tamil as a Natural Language processing—dictionary form of the words used in Lemmatizations. Stemming reduces the words, and Tamil is represented as a short sentence, comparing the differences in the Tamil dialect taken and represented. The methodology used is Clustering algorithms, which can group zonal dialects based on phonetic and semantic similarities using a Naïve Bayes classifier. We are using speech-to-text to identify the Tamil dialect. This zonal dialect is essential in entertainment, education, information, and business. More Exploration can be done using Zonal dialects in Classical Tamil. Machine learning plays a role in classification, Grouping, and Segmenting Natural Language processing. We have different dialects in the Single Language Tamil for a single word in Natural Language Processing. Encourages local people to communicate fluently in terms of transactions. Preserving local traditions and customs is one of the advantages of Zonal Dialects. It can be used in interviews, recordings, written and spoken texts, as well as debates. Linguistic Diversity, preservation of History, and cultural identity are significant concerns in Zonal dialects that use classical Tamil.***

Keywords: (must be 3-5), Zonal Dialects, Machine Learning, Naïve Bayes

Abbreviations:
NLP- Natural Language Processing
NLU: Natural Language Understanding
ASR: Automatic Speech Recognition
AST: Abstract Syntax Tree
TP: True Positives
TN: True Negatives
FP: False Positives
FN: False Negatives
CFG: Context-Free Grammars
ML: Machine Learning
SOV: Subject-Object-Verb

Manuscript received on 19 April 2025 | First Revised Manuscript received on 24 April 2025 | Second Revised Manuscript received on 04 May 2025 | Manuscript Accepted on 15 May 2025 | Manuscript published on 30 May 2025.
*Correspondence Author(s)
Perumal Sivaraman*, Department of Information Technology, UTAS - Nizwa, Oman (Nizwa), Oman. Email ID: sivasbc@gmail.com, ORCID ID: 0009-0003-5301-8968
Dr. Prabaharan G, Department of Computer Science and Engineering, Vel Tech Rangarajan, Dr. Sagunthala R and D Institute of Science and Technology, Morai (Tamil Nadu), India. Email ID: prabaharang@gmail.com, ORCID ID: 0000-0002-8365-5322
Dr. Senthil Kumar R, Department of Computer Science and Engineering, Jain Deemed University, Bengaluru (Karnataka), India. Email ID: senkr.raj@gmail.com, ORCID ID: 0000-0002-1393-756X

I. INTRODUCTION

NLP plays a crucial role in the Internet era, creating different lexicons for sentiment, emotion, and hate speech to improve the efficiency of the models [1].

An increase in robustness of translation and transliteration systems has also contributed to the rise of NLP Systems for Tamil text [2]. Spoken form is the one that people use in their daily language as they talk [3]. Computational Linguistics is used in machine translation, speech recognition, and text-to-speech synthesisers [4]. Dialects, colloquialisms, text-to-speech to extract from video footage [5]. Classical Tamil and its dialects are agglutinative, meaning they use affixes extensively to indicate grammatical relationships. Each word can encode tense, mood, aspect, person, and case, making morphological analysis critical. Tamil follows a Subject-Object-Verb (SOV) structure. Zonal dialects may exhibit deviations from this order due to phonetic simplifications or cultural influences. In classical texts, words often have multiple layers of meaning that depend on the context. Zonal dialects use simplified or metaphorical versions of classical expressions, which require semantic modelling. Tamil has a phonemic inventory that changes zonally. Pronunciation differences in dialects must be considered in speech and text processing. Classical Tamil often uses archaic letters and vocabulary forms that are uncommon in modern dialects.

Natural Language Understanding (NLU) is a branch of artificial intelligence that focuses on enabling machines to comprehend and interpret human language in a meaningful way. Linguistics is crucial in understanding human communication, cognition, and culture, making it a diverse and interdisciplinary field. Tamil is a Dravidian language predominantly spoken in the Indian state of Tamil Nadu and parts of Sri Lanka, Singapore, Malaysia, the United States, the United Kingdom, the Middle East, and among Tamil diaspora communities worldwide. Dialects are variations of a language that are specific to zones. They can differ in various aspects, including pronunciation, vocabulary, grammar, and usage. The primary advantage of understanding dialects is the ability to communicate effectively with people from different backgrounds. It will boost

trade and tourism globally. Tamil consists of the following letters.

Table-I: Tamil Letters

அஆஇஈஉஊஎஏஐஒஓஔ உயிரெழுத்துகள்;	பன்னிரண்டு
அஇஉஎஒ குறில் எழுத்துகள் (Sounds shorter)	
ஆஈஊஏஐஓஔ நெடில் எழுத்துகள் (Sounds Longer)	
வல்லினம்	க் ச் ட் த் ப் ற் (Sounds Hard)
மெல்லினம்	ங் ஞ் ண் ந் ம் ந் (Sounds Softer)
இடையினம்	ய் ர் ல் வ் ழ் ள் (Sounds Neither hard or soft and it is Intermediate)

The application of Identifying Zonal dialects consists of

a. Dialect-Specific Translation
b. Automatic Speech Recognition (ASR)
c. Sentiment Analysis in Dialects
d. Cultural Preservation
e. Organ donations
f. Movie Reviews Collection
g. GAudio
h. Hospitals
i. Insurance Sector

A. Challenges

i. *Scarcity of Annotated Data:* Limited corpora for Classical Tamil and its dialects.
ii. *Domain Adaptation:* The Natural Language Processing Axioms in Classical Tamil for Zonal Dialects using machine learning. Modern pre-trained models may not perform well on classical or dialectal texts without fine-tuning.
iii. *Phonetic and Orthographic Divergence:* The gap between written classical Tamil and spoken dialects complicates NLP tasks.

B. Future Directions

i. *Building Comprehensive Corpora:* Collecting and digitising classical Tamil literature and oral dialectal texts.
ii. *Multi-Dialect Pre-Trained Models:* Training models from scratch on Tamil-specific datasets for better contextual understanding.
iii. *Cross-Linguistic NLP:* This involves leveraging knowledge from other Dravidian languages to improve Tamil NLP systems. This fusion of linguistics and machine learning holds immense potential for preserving Tamil's linguistic heritage while making it accessible in the digital age.

C. Lexical Differences

NL understanding and Recognition. Natural Language Understanding (NLU) is a branch of artificial intelligence that focuses on enabling machines to comprehend and interpret human language in a meaningful way. It involves several key components:

i. *Syntax and Grammar:* Understanding the structure of sentences, including parts of speech and sentence construction.
ii. *Semantics:* Interpreting the meaning of words and phrases in context, which can involve recognizing synonyms, idioms, and nuances.
iii. தம்பி குதிக்காத - Brother, don't jump (In other regions/zones)
iv. **தம்பி** சாடாத - Brother, don't jump (In Kanyakumari and Nellai slang). In Google Translate, it provides the meaning of "**சாடாத**", which translates to "unspeakable".
v. *Pragmatics:* Understanding the intended meaning behind the words, considering context, tone, and social cues.
vi. நண்பன் - Friend (In Chennai)
vii. கூட்டுக்காரன் - Friend (In Kanyakumari and Nellai slang)
viii. தோஸ்து/ கூட்டாளி - Friend (In other regions). In Google Translate, it provides a collaborator for **கூட்டுக்காரன்**. Sentiment Analysis: Assessing the emotional tone behind a series of words to determine whether the sentiment is positive, negative, or neutral.

D. The Meaning of Slow is Expressed in Different Towns with Different Dialects

i. பைய - (Neutral with care) - (In Kanyakumari and Nellai slang)
ii. மெல்ல/மொல்ல/ மொள்ளமா - (Neutral with care)- (In Tiruvanamalai regions/zones)
iii. உஷாராக - (warning) - (In Tiruvanamalai regions/zones), கவனமாக - (warning) - (In Kanyakumari and Nellai slang), ஜாக்கிறதியாக - (warning) - (In other regions)
iv. *Contextual Understanding:* Grasping the broader context of conversations, including prior interactions and situational factors.

E. Morphology Top of Form

Morphology is the branch of linguistics that studies the structure and formation of words. It focuses on the internal structure of words and how they can be broken down into smaller, meaningful units called morphemes. Here are some key concepts related to morphology:

i. *Morphemes:* The smallest units of meaning in a language. They can be classified into:
ii. *Free Morphemes:* Stand alone as words (e.g.," கடலை")
iii. *Bound Morphemes:* Cannot stand alone and must attach to other morphemes (e.g.," வேர்க்கடலை")
iv. *Word Formation Processes:* Morphology examines how new words are created, which can involve the application of Natural Language Processing Axioms in classical Tamil for Zonal Dialects using machine learning.
v. *Derivation:* Creating a new word by adding prefixes or suffixes (e.g. "கடலை-> வேர்க்கடலை").
vi. *Inflexion:* Modifying a word to express grammatical features like tense,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

mood, number, or case (e.g.” வேர்க்கடலை சாப்பிடு”).

vii. *Compounding:* Combining two or more free morphemes to create a new word (e.g., "வேர்க்கடலைபர்பி ","வேர்க்கடலைமிட்டாய் ").

viii. *Morphological Rules:* Each language has specific rules governing how morphemes can be combined and altered, which can vary widely across languages.

Understanding morphology is crucial for linguistics, language education, and natural language processing, as it helps analyse and generate language.

Phonology is the branch of linguistics that studies the sound systems of languages, focusing on how sounds function and pattern in particular languages. Here are some key concepts related to phonology:

ix. *Phonemes:* The smallest units of sound that can distinguish meaning in a language. வேலை வேளை

- *Allophones:* Variations of a phoneme that occur in specific contexts but do not change meaning. For example, the /p/ in "pin" is pronounced with a puff of air (aspirated) while the /p/ in "spin" is not (unaspirated).
- *Syllable Structure:* Phonology examines how sounds combine to form syllables, typically consisting of a nucleus and may include an onset and a coda. (e.g.,”மா” vs “மாலை”) மா refers Mango and மாலை refers evening.
- *Stress and Intonation:* Phonology also studies stress and intonation patterns. In speech, words can affect meaning and grammatical structure. (e.g.,”அம்மா”)
- *Minimal Pairs:* Pairs of words that differ by only one phoneme, illustrating the phonemic distinctions in a language. (e.g.,”வேலை” vs “மாலை”)

F. Parser

A parser is a crucial component in natural language processing (NLP) and computer programming that analyses the structure of text or code. Its primary role is to break down input data into its constituent parts and understand the relationships between them. Here are the main aspects of parsing:

G. Types of Parsers

i. *Natural Language Parsers:* Used in NLP to analyse sentences, identifying grammatical structures (like subjects, verbs, and objects) and extracting meaning. They can be further categorised into:

ii. *Constituency Parsers:* Break down sentences into sub-phrases or constituents, showing how words group together.

iii. *Dependency Parsers:* Focus on the relationships between words, representing how words depend on each other.

iv. *Programming Language Parsers:* Analyse source code to check for syntax errors and create a representation of the code structure, often producing an abstract syntax tree (AST).

H. Parsing Techniques

i. *Top-Down Parsing:* It begins at the highest level of the parse tree and works downwards, trying to match the input with the grammar rules.

ii. *Bottom-Up Parsing*: This method starts with the input symbols and works upwards, combining them to form larger structures until it reaches the root of the tree.

iii. *Grammar:* Parsers rely on formal grammar, a set of rules that define how sentences or code are structured. Ordinary grammars include context-free grammars (CFG) and regular grammars.

I. Tokenization

Tokenisation is the process of splitting text into smaller units, called tokens, which can be words, phrases, or even individual characters, depending on the task and language model used. In natural language

The Natural Language Processing Axioms in Classical Tamil for Zonal Dialects Using Machine Learning. Processing (NLP), tokenisation is a critical pre-processing step because it structures raw text data into a format that algorithms can process effectively. Here's a breakdown of standard tokenisation techniques and their purposes:

i. *Word Tokenisation*

- *Definition:* Splits text into individual words.
- *Usage:* Common in tasks where each word's meaning and relationship to other words are essential, like in language models or text classification.
- *Example:* "The cat sat on the mat." becomes ["The", "cat", "sat", "on", "the", "mat", "."] “வெங்காயம்”,” விலை”,”குறைந்தது”

ii. *Subword Tokenization*

- *Definition*: Breaks text into subword units, helpful in handling out-of-vocabulary or rare words.
- *Usage*: Frequently used in Transformer models like BERT and GPT to improve flexibility in handling vocabulary.
- *Example*: "“வெங்காயம்” might be tokenized as [“வெங்”, “காயம்”].

iii. *Character Tokenisation*

- *Definition*: Splits text into individual characters.
- *Usage*: Useful for handling misspellings or languages with complex morphology and tasks involving creative text (e.g., poetry generation).
- *Example*: "“வெங்காயம்” becomes ["“வெ ", "ங் ", "கா ", "ய ", "ம் "].

iv. *Lexicon*

- *Lexicon* is a collection or database of words and their meanings, properties, and sometimes even relationships to other words. It functions much like a dictionary, but often includes additional details such as parts of speech, morphological forms, syntactic behaviour, and semantic relationships.

Here’s how a lexicon is applied and useful in various contexts:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

v. *General Linguistic Lexicon*

- A traditional linguistic
  A lexicon lists words, their definitions, usage, synonyms, and antonyms, helping users understand the structure and meaning of language.
- Often includes morphological details, like plural forms or conjugations, and phonological aspects, like pronunciation.

vi. *Computational Lexicon in NLP*

- *In NLP, a* computational lexicon stores definitions and properties applicable for algorithm processing. This includes:
- *Part of Speech (POS):* Identifying nouns, verbs, adjectives, etc., which helps with syntactic analysis.
- *Syntactic Rules:* Rules about how words combine with others, like verb-object relationships.
- *Semantic Information:* Concepts, senses, and meanings are key to understanding and generating natural language.
- *Domain-Specific Lexicons:* Custom lexicons tailored to specific fields, like medical, legal, or technical domains, where specialized vocabulary and definitions are essential.

Table-II: Dialect Morphology Spoken in Various Zones (Dialect-Tam)

S.No	Chennai	Kanyakumari	Nellai	Tiruvanamalai	Vellore	Kallakurichi	Villupuram	Pondicherry	English
AA	குதிக்காத	சாடாத	சாடாத	குதிக்காத	குதிக்காத	குதிக்காத	குதிக்காத	குதிக்காத	Don't Jump
AB	பிரச்சினை	களிப்பு	களிப்பு	பிரச்சினை	பிரச்சினை	பிரச்சினை	பிரச்சினை	பிரச்சினை	problem
AC	வெங்காயம்	பெல்லாரி	பெல்லாரி	வெங்காயம்	வெங்காயம்	வெங்காயம்	வெங்காயம்	வெங்காயம்	Onion
AD	நண்பன்	கூட்டுக்காரன்	கூட்டுக்காரன்	தோஸ்து/ கூட்டாளி	தோஸ்து/ கூட்டாளி	தோஸ்து/ கூட்டாளி	தோஸ்து/ கூட்டாளி	தோஸ்து/ கூட்டாளி	Friend
AE	நண்பா	மக்கா/மக்களே	நண்பா	நண்பா	நண்பா	நண்பா	நண்பா	நண்பா	Friend
AF	திருடன்	கள்ளன்	கள்ளன்	திருடன்	திருடன்	திருடன்	திருடன்	திருடன்	Thief
AG	மெல்ல/ மொல்ல	பைய	பைய	மெல்ல/ மொல்ல/ மொள்ளமா	மெல்ல/ மொல்ல	மெல்ல/ மொல்ல	மெல்ல/ மொல்ல	மெல்ல/ மொல்ல	Slow
AH	வேலை	சோலி	சோலி	வேலை	வேலை	வேலை	வேலை	வேலை	Job
AI	லுங்கி	சாரம்	சாரம்	லுங்கி	லுங்கி	லுங்கி	லுங்கி	லுங்கி	Dhoti casual
AJ	உடம்பு சரியில்லை	சீக்கு	சீக்கு	உடம்பு சரியில்லை	உடம்பு சரியில்லை	உடம்பு சரியில்லை	உடம்பு சரியில்லை	உடம்பு சரியில்லை	Disease
AK	சர்க்கரை	சீனி	சீனி	சர்க்கரை	சர்க்கரை	சர்க்கரை	சர்க்கரை	சர்க்கரை	sugar
AL	ஓரமா போ	தூர போ	தூர போ	ஓரமா போ	ஓரமா போ	ஓரமா போ	ஓரமா போ	ஓரமா போ	Go Aside
AM	லேய்	என்தே எந்த		டேய், என்னடா	டேய், என்னடா	டேய், என்னடா	டேய், என்னடா	டேய், என்னடா	what
AN	சும்மா இரு	சும்மா இரு	சும்மா இரு	கம்முன்னு இரு	கம்முன்னு இரு	கம்முன்னு இரு	கம்முன்னு இரு	கம்முன்னு இரு	Be silent
AO	கடைசில வந்து	கடைசில வந்து	கடைசில வந்து	அப்பாடா	அப்பாடா	அப்பாடா	அப்பாடா	அப்பாடா	Finally
AP	என்னலே / என்னடே	என்னலே / என்னடே	ஏலே	என்னம்மா என்னடா	என்னம்மா என்னடா	என்னம்மா என்னடா	என்னம்மா என்னடா	என்னம்மா என்னடா	What He/she
AQ	போறியா	நீ போறியா	நீ போறியா	நீ போறியா	நீ போறியா	நீ போறியா	நீ போறியா	நீ போறியா	Are you Going
AR	போறீங்களா	நீங்க போறீங்களா	நீங்க போறீங்களா	நீங்க போறீங்களா	நீங்க போறீங்களா	நீங்க போறீங்களா	நீங்க போறீங்களா	நீங்க போறீங்களா	Are you Going

									with respect
AS	நான் சொல்றேன்	நான் சொல்றேன்	நான் சொல்றேன்	நான் சொல்றேன்	நான் சொல்றேன்	நான் சொல் றேன்	நான் சொல்றேன்	நான் சொல்றேன்	I am telling you
AT	வேர்க்கடலை	வேர்க்கடலை	வேர்க்கடலை	வேர்க்கடலை/ மல்லாட்டை / மல்லாக் கோட்டை / நிலக்கடலை	வேர்க்கடலை / மல்லாட்டை / மல்லாக் கோட்டை/ நிலக்கடலை	வேர்க்க டலை / மல்லாட்டை / மல்லாக் கோட்டை /நிலக்க டலை	வேர்க்கடலை / மல்லாட்டை / மல்லாக் கோட்டை / நிலக்கடலை	வேர்க்கடலை மல்லாட்டை/ மல்லாக் கோட்டை / நிலக்கடலை	Groundnut
AU	கவனமாக	கவனமாக	கவனமாக	கவனமாக ஜாக்கிறதியாக	கவனமாக	கவனமாக	கவனமாக	கவனமாக	careful
AV	வேகமாக/ விரைவாக	சீக்கும்	சீக்கும்	விரைவாக	விரைவாக	விரை வாக	விரைவாக	விரைவாக	Speed

- Helps NLP systems perform better on domain-specific tasks, like medical diagnosis assistance or legal document analysis.

i. Sentiment Lexicons

- Sentiment lexicons are collections of words with associated sentiment values (e.g., positive, negative, neutral), valid in tasks like sentiment analysis.
- Examples include AFINN, Senti WordNet, and VADER lexicons, which tag words with polarity scores for sentiment-based NLP tasks.

Lexicons form the backbone of many NLP tasks by providing models with a reference for word meanings, properties, and interrelations, enhancing their ability to interpret and generate human-like text. The Natural Language Processing Axioms in Classical Tamil for Zonal Dialects using machine learning

ii. Lemmatization

Lemmatization is the process in natural language processing (NLP) of reducing words to their base or root form, called a lemma, which represents the "dictionary form" of a word.

Lemmatisation utilises vocabulary and morphological analysis to retrieve the correct base form, ensuring the root word is both grammatically accurate and semantically meaningful.

Key Aspects of Lemmatization

iii. Linguistic Precision

Lemmatization considers the word's part of speech (POS) (e.g., noun, verb, adjective) to return the accurate lemma.

For instance, "ஓடுகிறான் " becomes "ஓடு "

Tools for Lemmatization

Popular NLP libraries offer lemmatization tools, such as:

- **NLTK** (Python): Provides lemmatisation with the aid of WordNet.
- **SpaCy**: Provides efficient lemmatization with POS tagging.

iv. Stemming

Stemming is a natural language processing (NLP) process that reduces words to their root or base form, known as a *stem*, by removing prefixes and suffixes. Unlike lemmatization, which relies on a vocabulary to find a word's base form, stemming applies simple rules to strip affixes from words, often ignoring context and grammatical correctness.

Table-III: Zonal Dialects

Zonal Dialects	Difference in Dialects
Kanyakumari	18
Nellai	15
Tiruvanamalai	8
Vellore	8
Kallakurichi	8
Villupuram	8
Pondicherry	8

Key Aspects of Stemming

v. Rule-Based Reduction

Stemming uses a set of language-specific rules to strip suffixes (and sometimes prefixes) from words.

For example:

"running" becomes "run"

ஓடினான் becomes ஓடு

"happily," becomes "happi"

மகிழ்ச்சியுடன் becomes மகிழ்ச்சி

vi. Algorithm Simplicity

- Stemming algorithms are simpler and faster than lemmatization algorithms because they don't need to
- understand the context or grammar.
- They are typically heuristic-based and do not guarantee that the output is a real word (e.g., "Happiness" could become "Happi").

vii. Popular Stemming Algorithms Bottom of Form

- *Porter Stemmer:* One of the most used stemming algorithms, known for its balance between simplicity and effectiveness.
- *Lancaster Stemmer:* A more aggressive algorithm, often producing shorter stems, which can lead to more unusual or truncated results.
- *Snowball Stemmer:* An improved and multilingual version of the Porter Stemmer, offering more control and

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

consistency across languages.

viii. *Stemming vs. Lemmatization*

- *Stemming:* Uses rules to strip affixes quickly; does not ensure a proper dictionary word.
- *Lemmatization:* Uses vocabulary and POS tagging to return a correct base form, making it more accurate for many NLP tasks

II. MATERIALS AND METHODS

Naïve Bayes Classifier uses the Bayes' theorem to predict membership probabilities for each class, such as the probability that a given record or data point belongs to a particular class.

The MAP for a hypothesis with two events X and Y is

= max (P (X | Y))

= max (P (Y | X) * P (X))/P (Y)

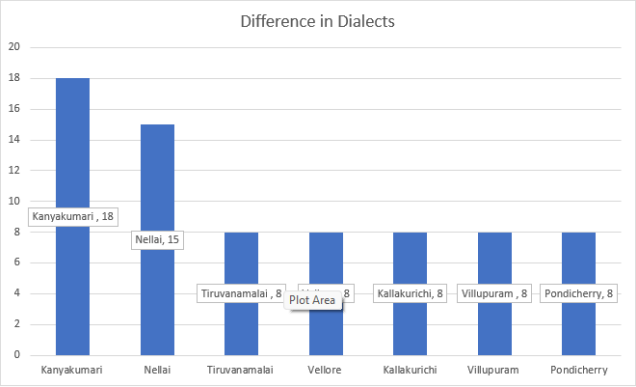
= max (P (Y | X) * P (X))

Here, P(Y) represents the probability of the evidence. It is used to normalize the result. It remains the same, so removing it would not affect the result.

Naïve Bayes Classifier assumes that all the features are unrelated to each other.

The presence or absence of a feature does not influence the presence or absence of any other feature.

There are 20 Chennai words, compared with the other Zones.



[Fig.1: Comparison of Differences in Dialects]

Kanyakumari and Nellai words differ from those of Chennai.

Kanyakumari and Nellai words differ from Tiruvannamalai, Vellore, Kallakurichi, and Villupuram. Among these, Kanyakumari ranks first in terms of the difference from Chennai in words.

Second, the Nellai words are different from the Chennai words.

Third, the words Tiruvannamalai, Vellore, Kallakurichi, Villupuram, and Pondicherry are different from Chennai words, but they are very close.

A. Implementation

Loading several Libraries (Linear algebra, pandas, matplotlib, pyplot and seaborn)

B. Dimensions of the dataset

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import os
data = '/ Dialect-Tam /.csv'
df = pd.read_csv(data, header=None, sep=',\s')
```

C. Confusion Matrix

i. *True Positives (TP):* True Positives occur when we predict that an observation belongs to a particular class and the observation belongs to that class.

ii. *True Negatives (TN):* True Negatives occur when we predict that an observation does not belong to a particular class and does not belong to that class.

iii. *False Positives (FP):* False Positives occur when we predict an observation belongs to a particular class but does not belong to that class. This type of error is called a Type I error.

iv. *False Negatives (FN):* False Negatives occur when we predict an observation does not belong to a particular class when the observation belongs to that class. This is a grave error, and it is referred to as a Type II error.

```
df.shape
df.head()
df.info()
```

D. Sklearn Dataset

```
from sklearn.datasets import Dialect-Tam
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
X, y = Dialect-Tam (return_X_y=True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=0)
gnb = GaussianNB()
y_pred = gnb.fit(X_train, y_train).predict(X_test)
print("Number of mislabeled points out of a total %d points : %d"  % (X_test.shape[0], (y_test != y_pred).sum()))
```

These dialects are spoken in various regions of Chennai, Kanyakumari, Nellai, Tiruvannamalai, Vellore, Kallakurichi, Villupuram and Pondicherry.

Understanding these dialects can be challenging, as various regions exhibit minor differences.

E. Representation Using Propositional Logic

F	G
குதிக்காத	சாடாத
பிரச்சினை	களிப்பு

Entailment

F ⊨ G

குதிக்காத ⊨ சாடாத

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In the formal definition model, if F is True, then G is also True

F ∨ G is true if and only if F or G is true

பிரச்சினை ∨ களிப்பு

F ∧ G is true if and only if both F and G are true

பிரச்சினை ∧ சாடாத

¬F It is true if F is false

F ≡ *G*It is true if and only if both F and G are true or both F and G are false

குதிக்காத ≡ சாடாத

The Natural Language Processing Axioms in classical Tamil for Zonal Dialects using machine learning

Propositional Theorem Proving

(F∧G) ≡ (G∧F) commutativity of ∧

(பிரச்சினை ∧ சாடாத) ≡ (சாடாத ∧ பிரச்சினை)

(F ∨G) ≡ (G∨F) commutativity of ∨

(பிரச்சினை ∨ களிப்பு) ≡ (களிப்பு ∨ பிரச்சினை)

¬(¬F) ≡ F double-negation elimination

வேலை¬(¬ வேலை) ≡ வேலை double-negation elimination

(F ⇒ G) ≡ (¬G ⇒¬F) contraposition

(வேலை ⇒ சோலி) ≡ (¬சோலி⇒¬ வேலை)

(F ⇒ G) ≡ (¬G∨F) implication elimination

(வேலை ⇒ சோலி) ≡ (¬சோலி ∨ வேலை)

(F ⇔ G) ≡ ((G ⇒ F) ∧ (F ⇒ G)) biconditional elimination

(வேலை ⇔ சோலி) ≡ ((சோலி ⇒ வேலை) ∧ (வேலை ⇒ சோலி))

¬ (F ∧G) ≡ (¬F∨¬G) De - Morgan

¬ (வேலை ∧ சோலி) ≡ (¬வேலை ∨¬ சோலி) De - Morgan

¬ (F ∨G) ≡ (¬F∧¬G) De - Morgan

¬ (வேலை ∨ சோலி) ≡ (¬வேலை ∧¬ சோலி) De - Morgan

Inference and proofs:

F=>G, F

G

வேலை => சோலி, வேலை

சோலி

Whenever the sentence வேலை is given, the sentence சோலி is inferred.

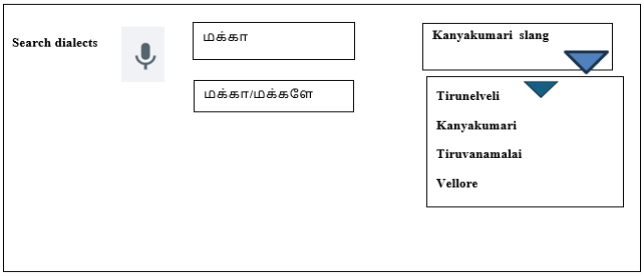
AND, Elimination, and the conjunctions can be inferred.

(F ∧ *G*)F can be inferred.

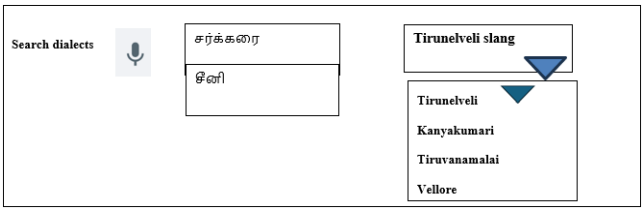
(வேலை ∧ சோலி) , வேலை can be inferred .

Table-IV: -Hardware Requirements

Hardware	Any computer with a large or higher, preferably cloud/or on-premises
OS	Windows/Linux/Macintosh
Processor	11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 1.69 GHz
RAM	16 GB
System Type	64-bit operating system, x64-based processor
Network Bandwidth	10 Gbps
Display Resolution	1920*1080
IOPS	16,000



[Fig.2: Dialects for சர்க்கரை with Different Slang]



[Fig.3: மக்கா Nellai, Kanyakumari Dialect]

Table-V: Software Requirements

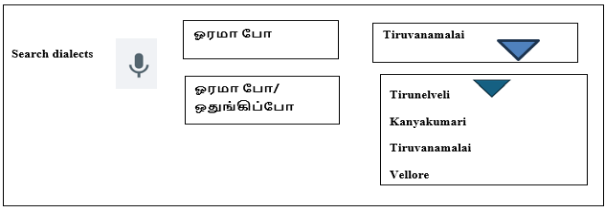
Operating System Windows (or) Linux
Python V3.7 / V3.8
PIP V23.1.2
Google Colab notebook, Latest version
Jupyter notebook, Latest version
pandas Latest version
numpy Latest version
matplotlib Latest version
seaborn Latest version

In Figure 3: மக்கா Nellai, Kanyakumari dialect

In Figure 2, the word " Chennai " (in Tamil, " சர்க்கரை is given as சீனி in Nellai, kanyakumari dialect, and it is given as சர்க்கரை in Tiruvanamalai, Vellore, Villupuram

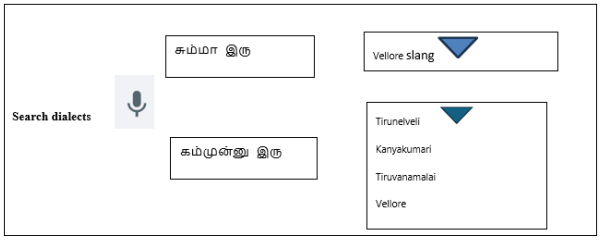
and in other zones. If we know these dialects, only we can understand the Language; otherwise, it is impossible. There are numerous dialects across different parts of the world. When we are taking different parts of the world, dialects will vary. It is an excellent challenge for learners to understand the word's exact meaning. A dedicated tool has been developed for Dialect prediction. It will be helpful in Computational Linguistics.

In Figure 3, Friend is referred to as நண்பா in Tiruvanamalai, Vellore, Villupuram, and other regions. If we are familiar with the dialects of Nellai and Kanyakumari, then we can understand the Language; otherwise, it is not possible. There are many dialects across the different parts of Natural Language. When considering other parts of the world, dialects vary. It is an excellent challenge for Learners to understand the exact word and meaning.



[Fig.4: Dialects for ஓரமா போ with Different Slang]

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)



[Fig.5: Dialects for சும்மா இரு with Different Slang]

In Figure 4: ஓரமா போ in Tiruvanamalai dialect is given as ஓரமா போ/ ஒதுங்கிப்போ in

Vellore, Villupuram and other zones. If we are aware of these dialects in Tiruvanamalai, then only we can understand the Language; otherwise, it is not

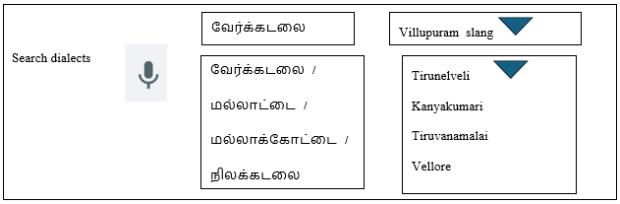
It is not possible. There are a lot of dialects across the different parts of the Natural Language for the word சும்மா இரு

In Figure 5: சும்மா இரு in vellore slang is given as கும்முன்னு இரு

Tiruvanamalai, Vellore, Villupuram and other zones. If we know these dialects in Vellore, only we can understand the language; otherwise, it is

not possible. There are many dialects across the various parts of the Natural Language.

When considering different parts of the world, dialects vary. It is an excellent challenge for learners to understand a word's exact meaning and dialect.



[Fig.6: Dialects for வேர்க்கடலை with Different Slang]

In Figure 6: Dialects for வேர்க்கடலை in Villupuram slang is given as வேர்க்கடலை, மல்லாட்டை, மல்லாக்கோட்டை or நிலக்கடலை in Tiruvanamalai, vellore, Tirunelveli, Kanyakumari and in other zones. If we are aware of these dialects, then only we can understand the Language completely

III. RESULTS AND DISCUSSIONS

Accuracy =TP + TN/TP+TN+FP+FN
Precision= TP/TP+FP
Recall =TP/TP+FN
F1-Score =2 (recall *precision/ recall +precision)

Creating a dataset for Dialects of conversation plays a vital role, and it is very challenging to list different dialects used in various zones and for various languages.

We classified using a Naïve Bayes classifier, which yielded an Accuracy of 97%.

The Natural Language Processing Axioms in classical Tamil for Zonal Dialects using machine learning

IV. CONCLUSIONS

In conclusion, the development of Zonal dialects is progressing steadily and is anticipated to usher in a new era marked by enhanced processing speeds. The proposed axioms in Classical Tamil for Zonal dialects demonstrate potential applicability across various natural languages. This approach offers high efficiency, making it a promising tool for future applications in natural language processing (NLP). The work serves as a foundation and eye-opener for further exploration across multiple languages. As research and development in natural language processing—primarily through machine learning—continue to evolve, monitoring advancements across different languages is vital. This will create new opportunities in diverse businesses, advertising, culture, and beyond.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/Competing Interests:** Based on my understanding, this article does not have any conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Mounikamarreddy, IIITH, India, Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP tasks in Telugu Language-ACM Transactions on Asian and Low-Resource Language Information Processing · April 2022. https://doi.org/10.1145/3531535
2. Omprakash Yadav1, Alcina Judy1, Praveen D'souza1, Calvin Galbaw1, Hinal Rane1-International Journal of Applied Sciences and Smart Technologies-2021. https://doi.org/10.24071/ijasst.v3i2.2826
3. Effect of Regional Dialects in Learning Tamil Language N. Sulochana- International Research Journal of Tamil-2022. http://doi.org/10.34256/irjt22s91
4. Dr. S. Suriya1, S. Nivetha2, P. Pavithran2, Ajay Venkat S.2, Sashwath K. G.2, Elakkiya G.2-Effective Tamil Character Recognition Using Supervised Machine Learning Algorithms- EAI Endorsed Transactions on e-Learning-2023. http://doi.org/10.4108/eetel.v8i2.3025
5. J. Angelin Jeba1, S. Rubin Bose2, R. Regin3, S. Suman Rajest4, Md Mahdi Hasan5-Intelligent Tamil Video Summarization: AI-Powered NLP, Translation, and Speech Integration for Enhanced Accessibility-FMDB Transactions on Sustainable Computer Letters-2024. https://doi.org/10.69888/FTSCL.2024.000179

AUTHOR'S PROFILE

Mr. Perumal Sivaraman works at the University of Technology, Applied Sciences - Nizwa, in Information Technology, and has a rich background in teaching and research. Over 25 years of teaching experience, coordinated with AICTE-funded training programs for various certifications and conducted FDPs and seminars.

He has published his research work in national and international journals, including those indexed in Web of Science, Scopus, and IEEE Explore. His areas of interest include Artificial Intelligence, Data Science, Machine Learning, and Federated Learning. Coordinated various activities like conferences, workshops, seminars, and other training programmes. Life member of the Indian Society of Technical Education (ISTE). Computer Society of India (CSI).

Dr. Prabaharan G has completed a PhD in Computer Science and Engineering**,** has rich teaching and research experience, and is currently associated with Veltech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology as an Associate Professor. He has over 20 years of teaching experience and has coordinated programs such as Faculty development programs, Seminars, and conferences. He has published his research work in various international and national journals, including those indexed in Web of Science, Scopus, and IEEE Explore.

Dr. Senthil Kumar has completed a PhD in Computer Science and Engineering, bringing a wealth of teaching and research experience. Over 25 years of teaching experience, and coordinated with industry-collaborated programs such as Oracle workforce development program, Infosys Campus Connect program, Wipro Mission 10X and AICTE-funded training programs for various certifications. He has published his research work in national and international journals, including those indexed in Web of Science, Scopus, and IEEE Explore. Published multiple patents for his research work and organised various national-level Faculty development programs, Workshops, conferences and seminars.