DEVELOPING A RETRIEVAL-BASED TAMIL LANGUAGE CHATBOT FOR CLOSED DOMAIN

Kumaran Kugathasan

198097X

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

August 2023

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 14/08/2023

The above candidate has carried out research for the Masters thesis/Dissertation

under my supervision.

Signature of the Supervisor: Date: 15/08/2023

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Uthayasanker Thayasivam, my supervisor, for his invaluable advice and guidance throughout this project. I am grateful for his willingness to make himself available whenever I needed his assistance. Without his support and encouragement, I could not have completed this project successfully.

Additionally, I would like to thank members of my progress review committee Prof. Sanath Jayasena and Dr. Charith Chitraranjan, for their insightful feedback and guidance, which were extremely helpful to me.

I am also grateful to the entire lecturers and research students at the DataSEARCH Research Centre for their assistance, feedback and the resources they provided to carry out the project.

Lastly, I would like to say thank you to my family and friends for their unfailing support.

ABSTRACT

Chatbots are conversational systems that interact with humans via natural language. Frequently, it is used to respond to user queries and provide them with the information they need. To build a highly functional chatbot, a good corpus and a variety of language-related resources are required. Since Tamil is a low-resource language those resources are not available for Tamil. Additionally, since Tamil is also a morphologically rich language, high inflexion and free word order pose key challenges to Tamil chatbots. Due to all the above reasons, it is evident that developing an effective End-to-End chat system is challenging even for a closed domain.

This study introduces a novel method for building a chatbot in Tamil by leveraging a dataset extracted from Tamil banking website's FAQ sections and extending it to encompass the language's morphological complexity and rich inflectional structure. Intent is assigned to each query, and a multiclass intent classifier is developed to classify user intent. The CNN-based classifier demonstrated the highest performance, achieving an accuracy of 98.72%.

While previous works on short-text classification in Tamil focused only on a few classes and used a very large dataset, our method produced a superior accuracy of over 98% using a small number of per-class examples even when there are 56 classes and additional challenges like class imbalance problem in the data. This shows our approach is better than any other approach for short text classification in Tamil.

The major contribution of this research is the generation of the first-ever chat dataset for Tamil. Our research is the first of its kind in Tamil to show how an efficient context-less chatbot can be built using short text classification. Although this project is done for the Tamil language and for the Banking domain, this approach can be applied to other low-resourced languages and domains as well.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FAQ	Frequently Asked Questions
NLU	Natural Language Understanding
NLG	Natural Language Generation
AIML	Artificial Intelligence Markup Language
POS	Part-of-speech
NER	Named-Entity Recognizer
BOW	Bag of Words
TF-IDF	Term frequency–inverse document frequency
SMOTE	Synthetic Minority Oversampling Technique

TABLE OF CONTENTS