Advanced Data Mining IA02

Multilingual Chatbot

(Deep Learning-based using Natural Language Processing)

By -
Sumeet Karani (16014022044),
Raunak Kankariya (16014022042),
Swapnil Ranadive (16014022067)
& Ketaki Mahajan (16014022050)

CONTENT

LITERATURE REVIEW

Title	Year of Publication	Methodology	Results	Research Gap
Multilingual Healthcare Chatbot Using Machine Learning	2021	NLP-based multilingual healthcare chatbot using five ML algorithms for disease prediction with English, Hindi, and Gujarati language support.	Random Forest achieved 98.43% accuracy; system demonstrated effective multilingual capabilities with TF-IDF and Cosine Similarity for query responses.	Most existing healthcare chatbots limited to English; need identified for multilingual solutions addressing rural Indian healthcare challenges.
Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: A multi-center study	2023	Developed DR-COVID, a multi-lingual AI chatbot on Telegram using: • Modified BERT + spaCy (weighted 0.8:0.2) for NLP • Google Translate API for multilingual support	• English Accuracy: 83.8% overall, 92.2% top-3 • Top Multilingual Accuracy: Portuguese (90%) • Faster & more accurate than WHO & NHS bots	• Existing bots: monolingual & triage-focused • Lacked support for open-ended, multilingual Q&A
Generative AI Agents, Build a Multilingual Chat GPT-based Customer Service Chatbot	2024	Contrasts traditional chatbot models with a ChatGPT-integrated architecture.	The ChatGPT-based chatbot showed superior performance, multilingual capabilities & reduced development effort.	Improving contextual understanding, addressing bias in responses, enhancing multilingual capabilities.
Multilingual Chatbot For Indian Languages	2023	Dual approach using MuRIL BERT for fixed-response and context-based QA with SQuAD fine-tuning.	MuRIL BERT outperformed other models with 76% accuracy.	Poor Indian language representation in existing models.

LITERATURE REVIEW

Title	Year of Publication	Methodology	Results	Research Gap
BilinBot: A Bilingual Chatbot using Deep Learning	2023	• Data collected in Bangla & English • Deep Learning with NLP: Tokenization, TF-IDF, Word2Vec, and BERT for feature extraction • RNN architectures using LSTM and GRU	• Outperformed state-of-the-art methods (better ROUGE scores) • High prediction accuracy for bilingual input	Insufficient resources for low-density (Bangla) language • Limited contextual dialogue handling and real-world deployment
Investigating Deep Learning for Predicting Multi-linguistic Interactions with a Chatterbot	2020	• Experimental RNN design with extensive model tuning (iterations, layers, cell counts) • Migration from TensorFlow 1 to TensorFlow 2 with Keras • Use of Word2Vec embeddings to boost prediction accuracy	• TF2-based models reached nearly 99% accuracy with improved convergence over traditional TF1 designs	• Tendency to reach local optima with complex datasets • Extended training times and hardware limitations affecting scalability
An Analysis of an Intelligent Chatbot Using Natural Language Processing	2022	Explores the development and deployment of a chatbot using Natural Language Processing (NLP). The paper reviews cloud-based chatbot technology and design strategies.	Highlights chatbot applications in academic and commercial environments. Discusses advantages of AI and NLP integration for enhancing user interaction.	The paper lacks deep exploration of advanced NLP techniques like deep learning in chatbots. There is no discussion on multilingual capabilities and advanced AI models in chatbots.

LITERATURE REVIEW

Title	Year of Publication	Methodology	Results	Research Gap
Natural Language Processing in Chatbots: A Review	2020	Reviews NLP techniques in chatbot development, such as tokenization, part-of-speech tagging, and sentiment analysis. Explores rule-based and machine learning approaches to NLP in chatbots.	Reviews applications of NLP across industries like customer service, information retrieval, and sentiment analysis. Discusses recent advances, including deep learning and neural language models.	The paper addresses challenges in NLP integration but lacks in-depth solutions for multilingual support and context handling. Limited exploration of real-world applications using advanced neural networks like transformers.

P01: Multilingual Healthcare Chatbot Using Machine Learning

➔ Provides multilingual healthcare assistance
➔ Combines Machine Learning and NLP for prediction & interaction

Motivation

- Lack of accessible medical advice in rural/multilingual areas
- Shortage of doctors, rising healthcare costs
- Need for intelligent, automated solutions

System Modules

Disease Prediction

- Dataset: 4920 records, 41 diseases, 17 symptoms
- Preprocessing: Binary symptom encoding using CountVectorizer
- Algorithms:
  - Random Forest (Best Accuracy: 98.43%)
  - KNN, SVM, Decision Tree, Multinomial Naive Bayes

Query Answering

- NLP pipeline: Tokenization, Stop-word removal, Stemming, TF-IDF vectorization and Cosine Similarity for matching FAQs
- Custom FAQ database

Technologies Used

- scikit-learn (ML models)
- speechrecognition (Voice Input)
- googletrans (Translation)
- gTTS, playsound (Voice Output)

Proposed Chatbot Architecture

Input:

- Voice/Text
- Multilingual (English, Hindi, Gujarati)

Processing:

- Translate input to English using googletrans
- Classify as disease/query
- Disease → ML model
- Query → FAQ NLP pipeline

Output:

- Translate back to original language
- Output via voice (gTTS) or text

Results

- Random Forest: Highest Accuracy (98.43%)
- Multilingual & multimodal support successful
- Improved accessibility in rural and regional settings

TABLE II. ALGORITHM EVALUATION METRIC RESULTS

Classification Algorithm	Accuracy	Precision	F1-Score
Random Forest	0.9843	0.9774	0.9781
Decision Tree	0.9712	0.9693	0.9697
SVM	0.9622	0.9547	0.9582
MNB	0.9539	0.9440	0.9454
KNN	0.9788	0.9731	0.9749

P02: Towards Deep Learning-Powered Chatbot for Translation Learning

➔ Development and Testing of a Multi-lingual NLP-based Deep Learning System in 10 Languages for COVID-19 Pandemic Crisis

DR-COVID Chatbot Development

- Telegram-based chatbot for COVID-19 FAQs
- Designed for both **monolingual** and **multilingual** users
- Answers **open-ended** COVID-19 questions using **deep learning**

Dataset

- Model trained on dataset with 2,728 English question–answer (Q&A) pairs sourced from **WHO**, **CDC**, **Singapore Ministry of Health**, and **medical guidelines.**
- Tested with 560 external questions in **9 languages**

Multilingual NLP Integration

Hybrid NLP stack:

- Modified **BERT** model (80% weight)
- **spaCy** NLP (20% weight)

Google Translate API used for multilingual support

Answers translated back into the user's language

Comparison with Other Systems

- **WHO Messenger** chatbot: **60%** accuracy
- **NHS Inform** chatbot: **50%** accuracy

Model Performance

- **English Performance**: Accuracy: 83.8%
- **Multilingual Accuracy**: Highest: Portuguese (90%)

TABLE 2 Multi-linguicism assessment for DR-COVID showing the top 3 performing non-English languages.

Language	*N*	Overall accuracy	Top 3 accuracy
Thai	20	0.600	0.700
Spanish	40	0.725	0.800
Portuguese	20	0.900	0.900

Limitations

- Lower accuracy on **unseen or multilingual questions**.
- **Translation errors** from Google Translate impact quality
- Lacks **contextual memory** (no multi-turn conversation)

P03: Generative AI Agents, Build a Multilingual Chat GPT-based Customer Service Chatbot

- ➔ implementation of a generative AI-based customer service chatbot using ChatGPT
- ➔ how ChatGPT's advanced natural language processing capabilities can be leveraged

Comparison of Conversational Models

- Traditional Conversational Models Limitations
- Generative AI Advantages

Natural Language Processing Techniques

- Intent Classification and Entity Recognition
- Self-supervised learning
- Transfer Learning using pre-trained models

Proposed Chatbot Architecture

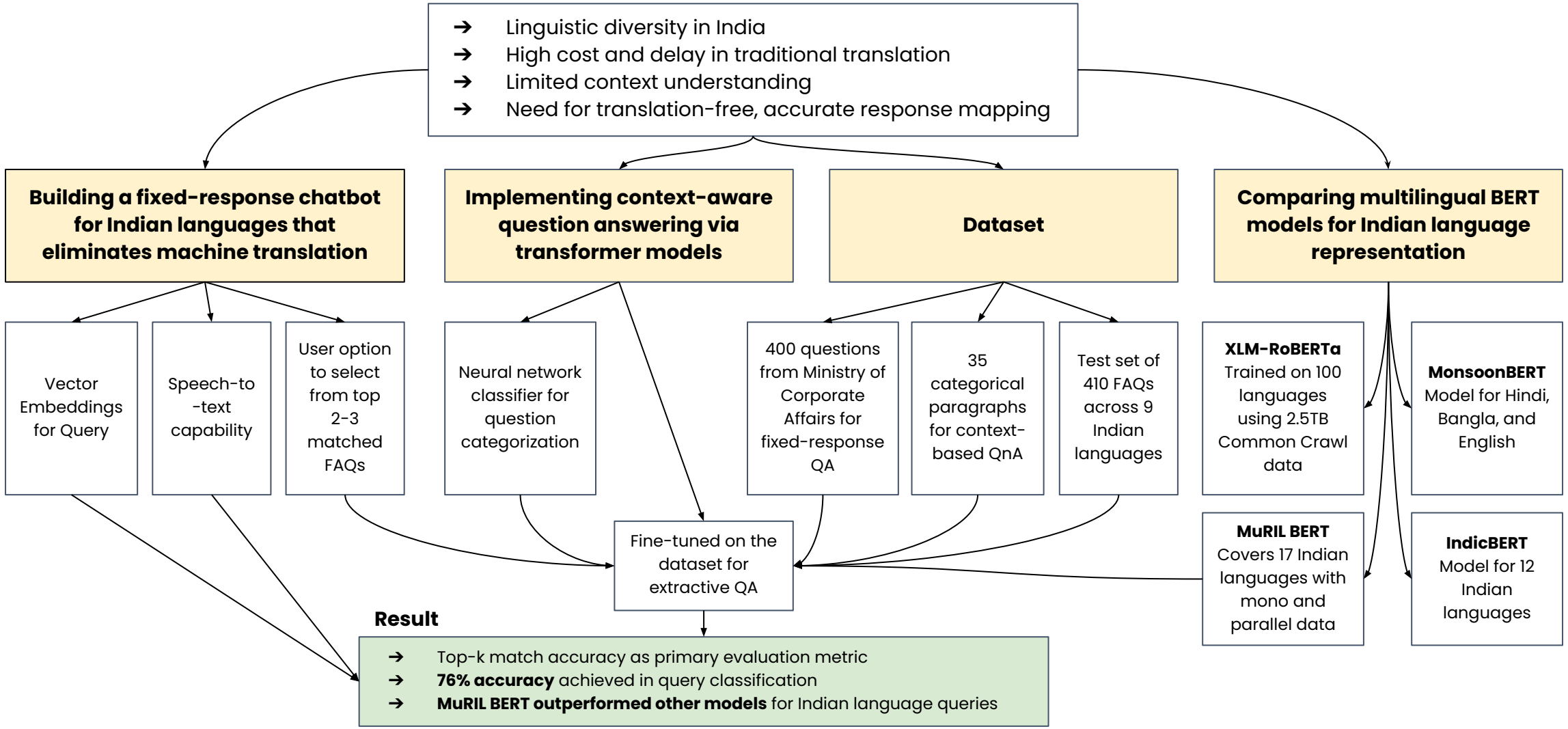
- Based on Generative Pre-trained Transformer model
- Contains 175 billion parameters across 96 layers
- Generates responses by predicting the next word based on previous sequences
- Implements Reinforcement Learning from Human Feedback (RLHF)
- Processes tokens across multiple languages

Key Findings and Results

- Superior Performance and Multilingual Capability
- Reduced Development Effort

Language	ChatGPT chatbot	Conventional chatbot
English	99%	75%
German	98%	09%
Mandarin	92%	03%
Norwegian	98%	08%
Korean	95%	01%

P04: Multilingual Chatbot For Indian Languages



P05: BilinBot: A Bilingual Chatbot using Deep Learning

- ➔ Development of a chatbot capable of conversing in both Bangla and English
- ➔ Leveraging deep learning and natural language processing to address the lack of bilingual conversational agents

Development of a Bilingual Chatbot (BilinBot)

- Methodology: Handles queries in Bangla & English
- Tools: Deep learning (RNN, LSTM, GRU) + NLP
- Motivation: Scarcity of Bangla NLP resources

Evaluation of Feature Engineering and Deep Learning Approaches

- Compared: Google BERT, TF-IDF, Word2Vec
- Result: BERT best for features, LSTM & GRU best for training
- Deep Learning: Tested RNN, LSTM, GRU

Creation and Preprocessing of Bilingual Datasets

- Dataset: English Q&A paraphrased & translated to Bangla
- Preprocessing: Keras Tokenizer, TF-IDF with n-grams
- Cleaning: Removed duplicates, corrected errors

Performance Benchmarking Against State-of-the-Art Chatbots

- Result: BilinBot outperformed other bilingual models
- Metrics: ROUGE-1, ROUGE-2, ROUGE-L

TABLE I. COMPARATIVE TEST SCORES (ENGLISH DATA).

Chatbot	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*
BilinBot	**0.96**	**0.96**	**0.96**	**0.94**	**0.94**	**0.94**	**0.96**	**0.96**	**0.96**
Conversational AI [12]	0.78	0.79	0.78	0.74	0.75	0.74	0.78	0.79	0.78
Seq2Seq AI [31]	0.72	0.73	0.72	0.68	0.69	0.68	0.72	0.73	0.72

TABLE II. COMPARATIVE TEST SCORES (BANGLA DATA).

Chatbot	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*
BilinBot	**0.76**	**0.76**	**0.76**	**0.73**	**0.73**	**0.73**	**0.76**	**0.76**	**0.76**
Conversational AI [12]	0.52	0.53	0.52	0.49	0.49	0.49	0.52	0.53	0.52
Seq2Seq AI [31]	0.48	0.49	0.48	0.46	0.46	0.46	0.48	0.49	0.48

P06: Investigating Deep Learning for Predicting Multi-linguistic Interactions with a Chatterbot

➔ Exploring deep learning architectures to predict and manage multilingual or code-switched conversations in chatbots, with a focus on adaptability and cross-linguistic comprehension.

Implementation of Multilingual Interaction Prediction Models

- Built predictive models for multilingual chatbot interactions using **sequence-to-sequence architecture**.
- Used deep learning models including **RNN** with **LSTM** and **GRU**.
- Tools included TensorFlow 1 & 2, **Keras**, and **Word2Vec** embeddings.
- Focused on next-turn prediction in multilingual or **code-switched conversations**.

Dataset Challenges and Solutions in Multilingual Contexts

- Used complex English text ("Declaration of Independence") and short phrases (like pangrams) to train and compare.
- Normalized and tokenized datasets using sequence padding and one-hot encoding.
- Observed that complex datasets hit local optima and required extensive tuning for convergence.

Performance Analysis Across Deep Learning Architectures

- Compared TensorFlow 1 (basic LSTM) and TensorFlow 2 (multi-layer LSTM with embedding).
- Word2Vec embeddings improved contextual understanding.
- Accuracy improved from 46.2% (TF1) to 99% (TF2 with Word2Vec).
- More layers and enhanced embeddings resulted in better generalization but required longer training.

Handling Code-Switching and Language Identification

- Tackled the challenge of Bangla-English mixed-language inputs.
- Used Word2Vec to create language-agnostic word vectors.
- Found that basic language ID and contextual representation improved the chatbot's relevance in responses.

Application in the BeatAI Educational Chatterbot

- Developed as a multilingual educational game where users "predict" or challenge the AI.
- Built as a web and Android application.
- Supported multiple languages (English, Russian, Polish, German, Spanish, Turkish, etc.).
- Integrated Yandex APIs for translation and language detection.

Limitations and Future Work

- Current limitations include local optima in model training and high resource usage.
- Lack of broader real-world testing beyond the game interface.

Future scope includes:

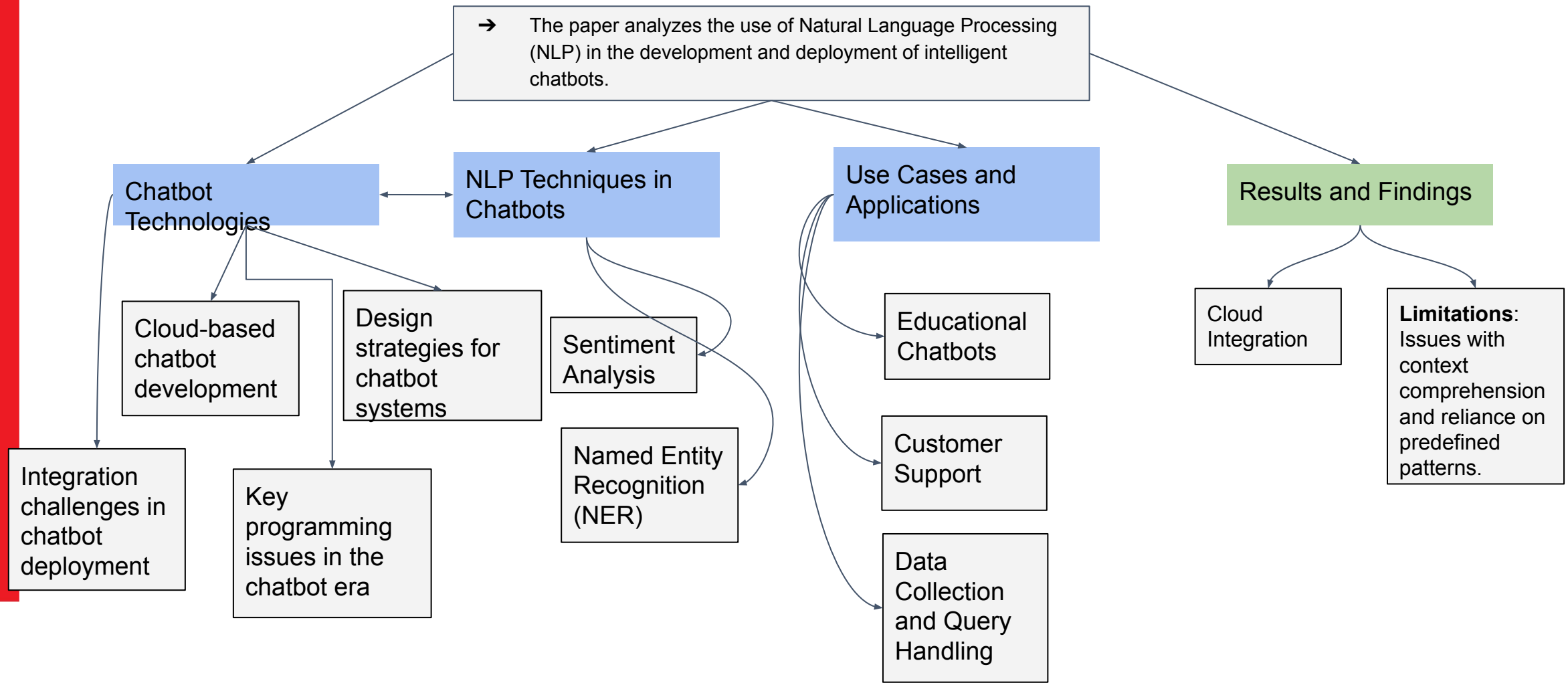
- Integration of Part-of-Speech tagging.
- Multilingual text auto-completion.
- Expansion to iOS platform and optimization for mobile hardware.

Results

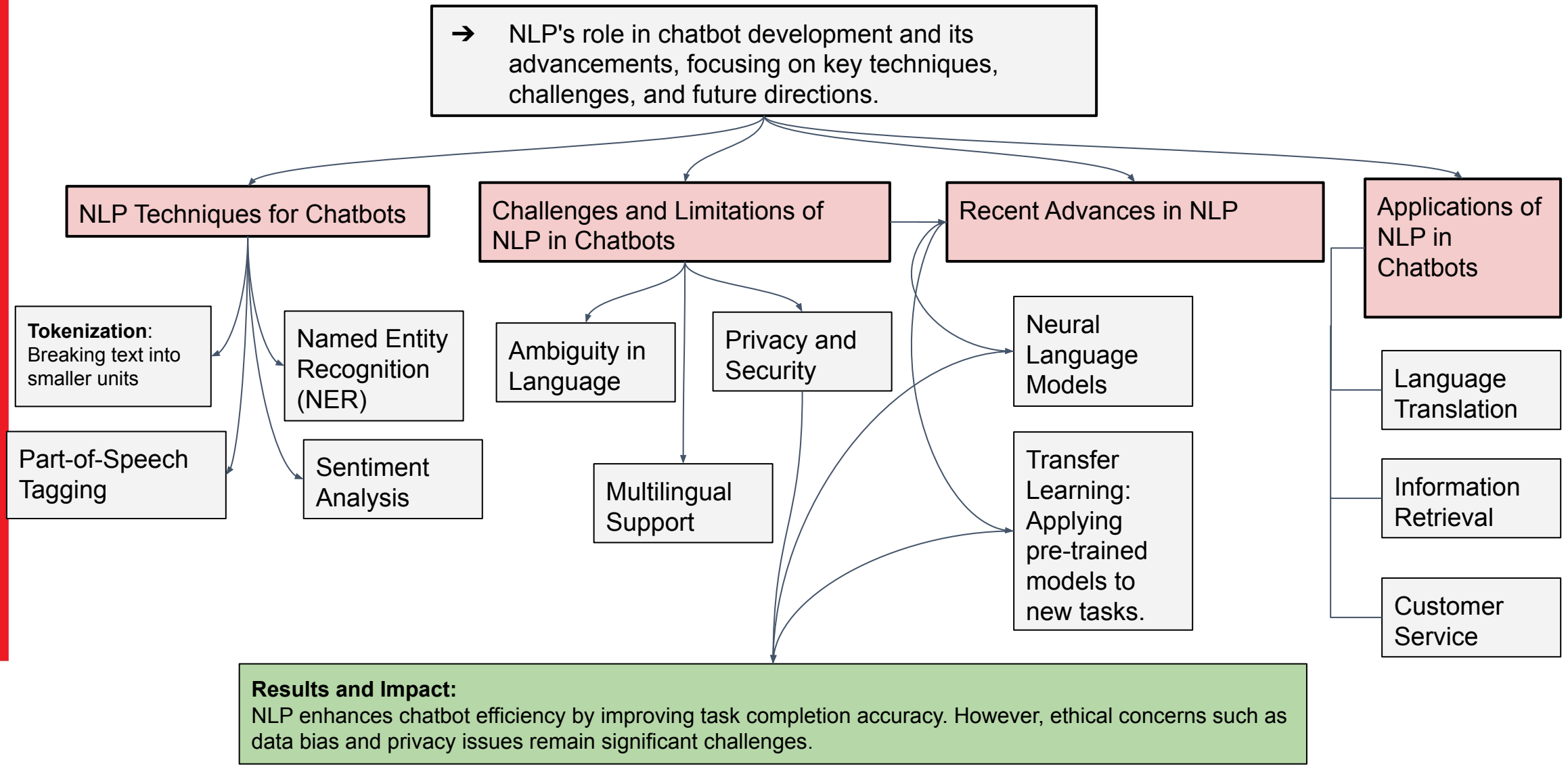
➔ The optimized **TensorFlow 2 model with Word2Vec** achieved up to **99% accuracy**, significantly outperforming the baseline TensorFlow 1 model (46.2%).

➔ Using **6-word input and multi-layer LSTMs** provided the best performance in multilingual prediction tasks.

P07: An Analysis of an Intelligent Chatbot Using Natural Language Processing



P08: Natural Language Processing in Chatbots: A Review



Conclusion

We studied multilingual chatbots using deep learning by analyzing research papers and building knowledge trees. This helped us understand key models and techniques used to handle multiple languages in conversation systems.

Benefits:

- Improved chatbot design for diverse languages
- Applications in global customer service, virtual assistants, and multilingual interfaces
- Enables scalable, intelligent systems with broader user reach