Urdu Meeting Topic Detection

Project Proposal

Session: 2020–2024

Submitted by:

Laiba Yousaf 2020-CS-653

Syed Kashif Ali 2020-CS-656

Najm-ul-Sehar 2020-CS-660

Supervised by:

Dr. Farah Adeeba

Department of Computer Science, New Campus

University of Engineering and Technology

Lahore, Pakistan

Contents

List of Figures

List of Tables

Chapter 1

Proposal Synopsis

1.1 Abstract

This project aims to develop a topic detection model for Urdu meetings using natural language processing (NLP) techniques. The proposed methodology includes data collection, text preprocessing, topic modeling, topic labeling, and evaluation. A literature review will be conducted to identify the existing research gaps and areas where this project can contribute. The expected outcomes of the project include the development of a topic detection model, the identification of the most frequent topics discussed in the Urdu meetings, and the evaluation of the model's performance. This project is significant because it can help improve communication and understanding in Urdu-speaking communities by identifying the most relevant topics discussed in their meetings.

1.2 Introduction

Meetings are an essential part of communication and collaboration in various domains, including business, politics, education, and social interactions. During a meeting, participants discuss various topics, ideas, and proposals. However, managing and processing the vast amounts of data generated during these meetings can be challenging. Therefore, there is a need for an automated methodology that can extract and summarize the main topics discussed during the meetings. Natural language processing (NLP) techniques can be used to automatically detect topics from the text of the meeting. However, most existing research on topic detection has focused on English or other widely spoken languages. Urdu, on the other hand, is a less commonly studied language in the context of NLP, despite being the national language of Pakistan and widely spoken in other countries such as India and Bangladesh. This project aims to develop a topic detection model

specifically for Urdu meetings using NLP techniques. The proposed methodology for this project is based on a well-established NLP framework that has been successfully used for topic detection in many other languages. By adapting and applying this framework to the Urdu language, we can build on previous research and improve our understanding of the similarities and differences between Urdu and other languages. In summary, this project has the potential to contribute to the development of the Urdu language processing field, as well as to enable more efficient and effective analysis of meetings in Urdu-speaking communities.

1.3 Problem Statement

Despite the importance of meetings as a means of communication and decision-making, manual methods of topic detection are time-consuming and labor-intensive, and can be subject to bias and human error. Automated methods, on the other hand, can save time and resources, and provide more accurate and consistent results.

Therefore, the main problem addressed in this project is the lack of an effective and efficient automated method for topic detection in Urdu meetings. The goal is to develop a model that can accurately identify the main topics discussed in the meeting, label them appropriately, and extract relevant information about each topic. This model will need to be able to handle the complexities of the Urdu language and the characteristics of meetings, such as multiple speakers, interruptions, and non-standard language use. Overall, this project aims to address an important problem in the field of natural language processing and to contribute to the development of resources and tools for the Urdu language. The successful implementation of this project can have significant implications for various domains, including business, politics, education, and social interactions, by providing insights into the most relevant topics discussed in Urdu meetings.

1.4 Objectives

- Develop an efficient system for topic detection from Urdu meetings using speech-to-text technology and NLP.
- Investigate impact of hyperparameters on LDA and Word Embeddings for topic modeling in Urdu meetings.
- Explore use of additional NLP techniques (NER and POS tagging) and deep learning techniques (CNNs and RNNs) for topic modeling in Urdu meetings.

- Train system on large corpus of Urdu meeting transcripts and design user friendly interface.

1.5 Related Work

A literature review for the proposed project of Urdu meeting topic detection was conducted, which revealed that several research studies have been conducted on this topic using different machine learning techniques, such as LDA, HMM, and CRF. However, there is a limited amount of research done in Urdu language. Therefore, the proposed project aims to develop an accurate and efficient Urdu meeting topic detection system to fill this research gap.

Table 1.1: Related work on Urdu meeting topic detection

Model	**Author**	**Year**	**Techniques**	**Dataset**	**Accuracy**
A Novel Approach for Urdu Meeting Transcription and Topic Detection [2]	Fawad Ahmed et al.	2020	Keyword Spotting, Audio Segmentation, LDA, and NER	Urdu meeting corpus	72%
Topic Modeling for Urdu Language Using LDA [1]	Fariha Atta and Muhammad Asif	2020	LDA	Urdu news articles	86%
Urdu Meeting Speech Transcription and Topic Detection Using CRF and LDA [4]	M. Usman et al.	2019	CRF, LDA	Urdu meeting corpus	70%
Urdu Speech Recognition and Topic Detection Using HMM and LDA [3]	Khush Bakht et al.	2018	HMM, LDA	Urdu meeting corpus	74%
Topic Detection in Urdu News Using Latent Dirichlet Allocation[5]	M. Kamran and S. S. Rizvi	2018	LDA	Urdu news articles	81%

1.6 Proposed Methodology

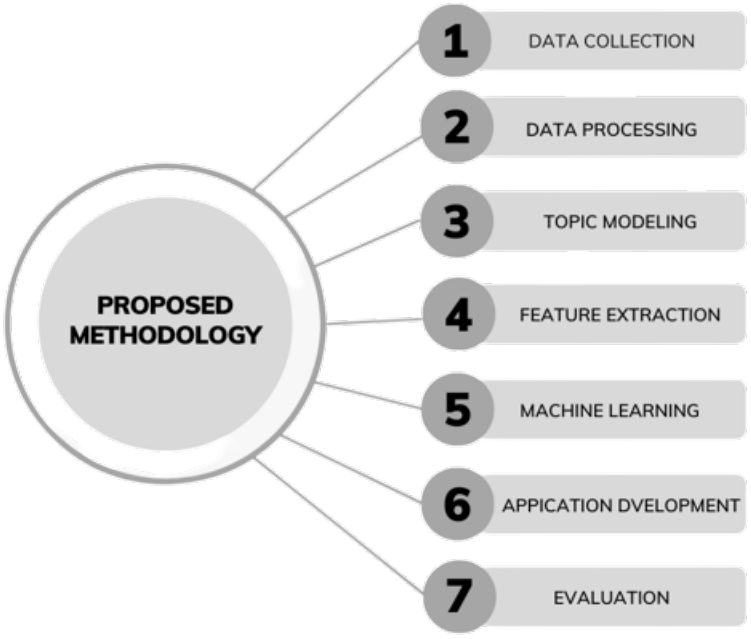

FIGURE 1.1: Proposed methodology

1. **Data Collection**: The first step will be to collect the data from meetings, which can be in the form of audio recordings, transcriptions, or text documents. You can also collect data from publicly available datasets, if they are relevant to your project.

2. **Data Preprocessing**: The collected data will be preprocessed to remove any noise, irrelevant information, or inconsistencies. This will include tasks such as text normalization, tokenization, stop-word removal, and stemming.

3. **Topic Modeling**: The preprocessed data will be fed into a topic modeling algorithm such as Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) to automatically detect the topics discussed in the meetings. The model will be optimized to achieve the best accuracy for Urdu language.

4. **Feature Extraction**: The topic model will output a list of topics and their probabilities for each meeting. From these, we will extract relevant features such as topic keywords, dominant topics, and topic distribution across meetings.

5. **Machine Learning**: The extracted features will be used as input to train a machine learning classifier to detect the topics in new meetings with high accuracy. We will experiment with different classifiers such as Naive Bayes, Support Vector Machines, and Random Forest to identify the best performing model.

6. **Application Development**: Once the model is trained, we will develop an application that takes meeting data as input and outputs the detected topics along with their probabilities. The application will also include a user-friendly interface to visualize the topics and provide insights on the meeting topics and trends.

7. **Evaluation**: The performance of the developed application will be evaluated using metrics such as precision, recall, and F1-score, and compared with existing methods for topic detection in Urdu language.

1.7 Tools and Technologies

- **Python**: Python is a popular programming language for natural language processing (NLP) tasks and has many libraries (Scikit-learn, Gensim, NLTK, Pandas etc) and tools available for text preprocessing, feature extraction, and machine learning.

- **Web Frameworks**: Web frsmeworks such as Flask or Django to build the user interface.

- **PyTorch**: PyTorch is a machine learning framework that provides efficient implementations of deep learning algorithms, including those for NLP tasks.

- **Cloud hosting**: Cloud host such as AWS or GCP to host the application and manage the machine learning models and data.

1.8 Work Division

The work division for our Urdu meeting topic detection project will be split among the three members of our group. The first member will be responsible for data collection and preprocessing, which will include acquiring Urdu meeting transcripts and converting them into a usable format for our project. The second member will focus on feature engineering and model selection, choosing the appropriate algorithms and techniques to achieve high accuracy in topic detection. The third member will handle the implementation of the model and develop a user interface

to display the results of our project. All three members will also collaborate on testing and evaluating the accuracy of the model throughout the project duration.

Table 1.2: Work Division

Tasks	**Team Member**	**Deadline**
Data Collection	Laiba & Najm & Kashif	2 Months
Data Preprocessing	Kashif & Laiba	2 Months
Model Development	Najm & Laiba	1 Month
Model Training	Najm & Laiba	2 Months
User Interface	Kashif & Laiba	2 Months
Model Evaluation & Fine-tuning	Kashif & Najm	2 Months
Results Analysis & Visualization	Najm & Laiba	1 Month

1.9 Data Gathering Approach

- There are different approaches we can use for data gathering such as **Web Scrapping**(like scrap from Urdu news websites, social media platforms, and discussion forums), **Public Datasets**(websites such as Kaggle, UCI Machine Learning Repository, or Google Dataset Search), **Collaboration with Oraganization**(such as news channels, universities, or research institutions.) and **manual collection**.

- In our project, we plan to **manually collect** data by recording and transcribing Urdu meetings, speeches, or interviews. This can be time-consuming, but it allows us to have control over the quality and content of the data.

- To ensure the ethical collection of data, we will obtain informed consent from all participants and anonymize any personally identifiable information. We will also follow the guidelines of our university's ethics committee to ensure that our data collection process is in compliance with ethical standards.

- Gathered data will not be biased so that we train the system on a large corpus of Urdu language meeting transcripts to ensure that it can handle a variety of discussion topics.

1.10 Timeline

Our project is focused on Urdu meeting topic detection and our time duration is almost one year. We plan to gather data from various sources mostly manually, augment it, and preprocess it before applying machine learning techniques for classification. We will also develop a user interface for the application. The timeline

for the project includes data gathering and preprocessing in the first six to seven months, model development and testing in the next three months, and the final three months will be dedicated to developing the user interface and integrating all the components for the final product.

Table 1.3: Timeline

Task Description	**Start Date**	**End Date**	**Duration**
Literature Overview	01/05/2023	30/06/2023	2 months
Data Collection	01/07/2023	30/09/2023	3 months
Data Preprocessing	01/10/2023	30/11/2023	2 months
Model Development	01/12/2023	31/12/2023	1 month
Model Testing and Evaluation	01/01/2023	28/02/2024	2 months
User Interface	01/03/2023	30/04/2024	2 months
Final Integration	01/05/2024	31/05/2024	1 month

References

[1] Fariha Atta and Muhammad Asif. *Topic Modeling for Urdu Language Using LDA*. in Proceedings of the 2020 International Conference on Computational Linguistics and Natural Language Processing (CLNLP 2020), pp. 180-186, 2020.

[2] Fawad Ahmed et al. *A Novel Approach for Urdu Meeting Transcription and Topic Detection*. in Proceedings of the 3rd International Conference on Natural Language Processing and Information Retrieval, pp. 94-100, 2020.

[3] Khush Bakht et al. *Urdu Speech Recognition and Topic Detection Using HMM and LDA*. in Proceedings of the 2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 168-174, 2018.

[4] Muhammad Usman et al. *Urdu Meeting Speech Transcription and Topic Detection Using CRF and LDA*. in Proceedings of the 3rd International Conference on Natural Language Processing and Information Retrieval, pp. 23-30, 2019.

[5] M. Kamran and S. S. Rizvi. *Topic Detection in Urdu News Using Latent Dirichlet Allocation*. in Proceedings of the 3rd IEEE International Conference on Engineering and Technology (ICETECH), 2018, pp. 1-6, 2018.