Large Scale Textual Analysis for Multilingual Semantic Search

A Thesis Submitted

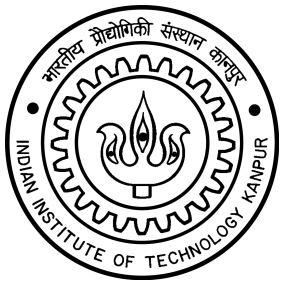
In Partial Fulfillment of the Requirements

For the Degree of Master of Technology

by

Pranshu Sahijwani

21111048



to the

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

May, 2023

Page intentionally left blank

Declaration

This is to certify that the thesis titled **"Large Scale Textual Analysis for Multilingual Semantic Search"** has been written by me. It describes the research I conducted under the guidance of **Prof. Purushottam Kar and Prof. Nisheeth Srivastava**.

It is, to the best of my knowledge, an original work, both in terms of research content and narrative, and has not been submitted elsewhere, either in part or in its entirety, for a degree. In addition, the relevant state of the art and collaborations (if any) have been appropriately acknowledged with citations and attributions, in accordance with established norms and procedures.

Name: Pranshu Sahijwani (21111048)
Program: Master of Technology
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur, Kanpur, 208016
May, 2023

Page intentionally left blank

Declaration (To be submitted at DOAA Office)

I hereby declare that

1. The research work presented in the thesis titled **"Large Scale Textual Analysis for Multilingual Semantic Search"** has been conducted by me under the guidance by my supervisor(s) **Prof. Purushottam Kar and Prof. Nisheeth Srivastava**.

2. The thesis has been formatted as per Institute guidelines.

3. The content of the thesis (text, illustration, data, plots, pictures etc.) is original and is the outcome of my research work. Any relevant material taken from the open literature has been referred and cited, as per established ethical norms and practices.

4. All collaborations and critiques that have contributed to giving the thesis its final shape have been duly acknowledged and credited.

5. Care has been taken to give due credit to the state-of-the-art in the thesis research area.

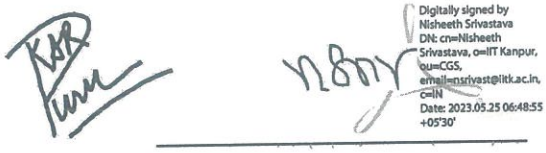
6. I fully understand that in case the thesis is found to be unoriginal or plagiarized, the Institute reserves the right to withdraw the thesis from its archive and also revoke the associated degree conferred. Additionally, the Institute also reserves the right to apprise all concerned sections of society of the matter, for their information and necessary action (if any).

Name: Pranshu Sahijwani
Program: Master of Technology
Department of Computer Science and Engineering
Roll No.: 21111048
Indian Institute of Technology Kanpur, Kanpur, 208016
May, 2023

Page intentionally left blank

Certificate

The work in the thesis titled **"Large Scale Textual Analysis for Multilingual Semantic Search"** by **Pranshu Sahijwani** was completed under my supervision and has not been submitted elsewhere for a degree.



Prof. Purushottam Kar and Prof. Nisheeth Srivastava
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
Kanpur, 208016
May, 2023

Page intentionally left blank

Abstract

Name of student: **Pranshu Sahijwani** Roll no: **21111048**
Degree for which submitted: **Master of Technology**
Department: **Department of Computer Science and Engineering**
Thesis title: **Large Scale Textual Analysis for Multilingual Semantic Search**

Name of thesis supervisor: **Prof. Purushottam Kar and Prof. Nisheeth Srivastava**
Month and year of thesis submission: **May, 2023**

This thesis introduces a comprehensive approach for language detection, linguistic analysis, and grievance clustering in large datasets. The first part focuses on language detection, employing Unicode character ranges and clustering techniques to accurately identify languages such as Hindi, English, Punjabi, Gujarati, Telugu, Tamil, Kannada, Odia, and Bengali. In the next section, dynamic programming is utilized to determine the most probable segmentation of input text, while an HNSW index facilitates the creation of a nearest neighbor dictionary. Spelling correction and word mapping techniques are explored in subsequent sections.

Additionally, a method for clustering similar grievances using text embeddings and the HNSW index is proposed to address the issue of bulk campaigns in large datasets. Finally, the thesis presents a system for language detection, text cleaning, translation, and similarity search, enhancing multilingual data analysis and aiding language processing and information retrieval. Overall, these contributions provide valuable tools for various applications in linguistic analysis and data processing.

Page intentionally left blank

Acknowledgments

I am extremely grateful to have had the privilege of being guided by Prof. Purushottam Kar and Prof. Nisheeth Srivastava from the Department of Computer Science and Engineering. Their unwavering support, valuable insights, and constant guidance have been instrumental in shaping the outcome of this thesis.

Prof. Kar's depth of knowledge, meticulous attention to detail, and insightful feedback have been instrumental in refining my research approach. His expertise in the field of computer science has provided me with a solid foundation and a broader perspective on the subject matter. I am truly indebted to him for his invaluable guidance and mentorship.

I would also like to express my heartfelt appreciation to Prof. Srivastava for his continuous encouragement and valuable inputs throughout the thesis work. His profound understanding of the subject matter, constructive criticism, and timely guidance have been instrumental in enhancing the quality and rigor of my research. I am immensely grateful for his support and belief in my abilities.

I extend my deepest gratitude to Prof. Purushottam Kar and Prof. Nisheeth Srivastava for their unwavering commitment, mentorship, and academic guidance. Without their invaluable support, this thesis would not have been possible.

Additionally, I am indebted to my family and friends(Archi, Tabish, Abhishek, Nimit and Manish) for their unwavering encouragement and understanding throughout this undertaking. Their love, patience, and belief in my abilities have been instrumental in my personal and academic development.

I am especially grateful to Gaurav Tank from MSR, IIT Kanpur for his invaluable assistance in the redaction of personally identifiable information (PII) in the Gujarati text within the results section of Chapter 4. His expertise and dedication significantly enhanced the quality and integrity of this research.

I would like to express my sincere gratitude to Olivier Commowick for generously providing the thesis template that served as the foundation for this work. The template, available at http://olivier.commowick.org/thesis_template.php, proved to be an invaluable resource in structuring and organizing my thesis according to the requirements of the Indian Institute of Technology Kanpur.

Thank you to everyone who has been a part of this incredible journey. Your contributions have left an indelible mark on this thesis, and I am truly grateful for your support.

Pranshu Sahijwani
May, 2023

Contents

List of Figures

List of Tables

List of Algorithms

Page intentionally left blank

Chapter 1

Language Distribution Analysis

Contents

Abstract *In this chapter, a concise methodological approach for detecting the language of text using Unicode character ranges and clustering techniques is presented. The language ranges for each language of interest, including Hindi, English, Punjabi, Gujarati, Telugu, Tamil, Kannada, Odia, and Bengali, are first defined. Next, a function is applied to count the number of characters in each language range for a given sentence, and the language with the highest count is determined. Furthermore, a clustering algorithm is used to classify text as either English or a mixture of Hindi and Latin characters. A reliable and efficient way to automatically detect the language of text is provided by the approach, which can be valuable in a variety of applications.*

1.1 Introduction

In the interconnected world of today, e-government applications are gaining popularity. However, language barriers can frequently impede effective citizen-government communication. In response to this difficulty, language distribution analysis has emerged as a viable method for enabling cross-language semantic queries. This chapter provides a tally of the language distribution of complaints submitted to e-government applications. By identifying the language of the submitted complaints, this method enables efficient data processing and analysis. This study's ultimate objective is to improve the accessibility and efficacy of e-government applications for citizens with diverse linguistic backgrounds.

1.2 Problem Statement

Perform language distribution analysis on a vast collection of grievances received from e-governance applications to identify the prevalence of different languages used by the citizens while filing complaints.

1.3 Related Work

1. M. Chen and K. Zhou, "Detecting Language of Web Pages using Unicode Character Distribution," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Hong Kong, China, 2006, pp. 617-620, doi: 10.1109/WI-IAT.2006.126.

   This paper is related to the first step in the work described, where Unicode ranges are defined for different languages. The authors propose a method for detecting the language of web pages using Unicode character distribution. They use a set of character frequency features based on Unicode character ranges to identify the language of the page.

2. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

   This book chapter discusses the task of language identification and describes several methods for achieving it, including the use of character n-grams and language-specific dictionaries. The second step in the work described, creating a function to detect language, could draw inspiration from some of these methods.

3. T. N. Sainath and A. Mohamed, "Automatic Language Identification Using Deep Neural Networks," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, 2015, pp. 4914-4918, doi: 10.1109/ICASSP.2015.7178947.

   This paper is relevant to the third step in the work described, where the language detection function is applied to a dataset. The authors propose a deep neural network-based method for automatic language identification, which achieves state-of-the-art performance on several benchmark datasets.

4. A. Anand, S. K. Bhatia, and S. Saha, "Clustering of Indian Language Text Documents using K-Means Algorithm," in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2017, pp. 1636-1640, doi: 10.1109/RTEICT.2017.8256676.

   This paper is related to the fourth step in the work described, where clustering is performed on non-Roman script sentences using the KMeans algorithm. The authors apply KMeans to cluster Indian language text documents and evaluate the performance of different feature extraction methods and clustering parameters.

5. J. Gao and Y. Fan, "A Scalable and Accurate Language Identification System for Large-scale Texts," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, pp. 3086-3095, doi: 10.18653/v1/D18-1343.

   This paper describes a system for language identification that is scalable to large-scale texts. The authors use a character-based convolutional neural network (CNN) to learn language-specific features and achieve high accuracy on several benchmark datasets. The approach could potentially be used to improve the language detection function in the work described.

1.4 Methodology

1.4.1 Preprocessing

The purpose of these pre-processing tasks is to clean and prepare the text data for further analysis. Here is a detailed methodological approach:

1. **Text Cleaning** The function accepts a text string as input and applies a series of text cleansing operations to it. The function specifically converts the text to lowercase, removes all punctuation, any digits, any excess white space, and any special characters. Finally, the cleaned content is returned.

2. **Text Frequency Analysis: Histogram of Alphabet Letters** The function takes in a string of text as input and computes a histogram of the frequency of each letter in the alphabet. The function returns an array of length 26, where each element corresponds to the frequency of a specific letter in the alphabet. The array is normalized using numpy's linalg.norm function before it is returned.

3. **Iterate over the data** The script then iterates over all the CSV files in a directory. For each CSV file, the script reads it into a pandas dataframe using the 'read_csv' function. The dataframe is then filtered to only include the columns 'reg_no' and 'subject_content'. The column names are renamed to more descriptive names.

4. **Clean the text data** The 'clean_text' function is applied to the 'subject_content' column of the dataframe using the 'apply' function. The cleaned text is stored in a new column named 'subject_content_cleaned'.

5. **Compute histograms for each row** The 'freq' function is applied to the 'subject_content_cleaned' column of the dataframe using the 'apply' function. The resulting histograms are stored in a new column named 'histograms'.

6. **Save preprocessed data:** Finally, the preprocessed dataframe is saved as a pickle file. The filename is constructed using the original filename, but with the extension replaced with '.pkl'.

1.4.2 Core functionality

1. **Defining Unicode ranges for various languages** We define Unicode ranges for different languages, which include characters unique to each language. These ranges help us identify the language of a given sentence by counting the number of characters in each range.

2. **Creating a function to detect language** We create a function that takes a sentence as input and detects the language by counting the number of characters in each language range. The function returns the language of the sentence as output.

3. **Detecting language of each sentence:** We use the language detection function to identify the language of each sentence in the dataset. We then split the dataset into two parts based on whether the sentences are in Roman script or not.

4. **Clustering Roman script sentences** We perform clustering on the Roman script sentences using the KMeans algorithm to classify them into two groups: English and Hindi-Latin. The KMeans algorithm is a type of unsupervised machine learning algorithm used for clustering data points.

5. **Counting documents in each language group:** Finally, we count the number of documents in each language group and save the counts to a pickle file. This allows us to easily access the counts later on for further analysis or visualization.

1.5 Experiments

1.5.1 Cluster Reassignment of Grievances using Word Frequency Analysis

1. Data Collection: The code reads data from a text file containing 100000 grievances and two CSV files containing grievances from two clusters built using character frequency histograms.

2. Data Preprocessing: The text data is preprocessed by removing links, special characters, and numbers using regular expressions. The text is then converted to lowercase and split into words using the ‘split‘ function.

3. Feature Extraction: The frequency distribution of words in the text is calculated using the ‘FreqDist‘ function from the ‘nltk‘ library.

4. Clustering: The words are clustered into two groups using the clusters calculated earlier using KMeans and character frequency histogram approach

5. Score Calculation: The code calculates the scores for English and Hindi at both the word level and the sentence level.

   (a) At the word level, the scores for each word are calculated as the percentage of occurrences of each word in each cluster. This is done by creating a list

of frequency distributions for each cluster using the FreqDist function from the nltk library. The frequency distributions are then used to calculate the percentage of occurrences of each word in each cluster and store these values in a dictionary called perc_dict.

(b) At the sentence level, the average scores for English and Hindi for each sentence are calculated using the build function. This function takes a list of words as input and returns a tuple containing the average scores for English and Hindi. The function calculates these scores by summing up the scores for each word in the input list (as stored in perc_dict) and dividing by the total number of words in the list.

6. Cluster Reassignment: The clusters are reassigned based on these average scores by comparing the scores for English and Hindi for each cluster and reassigning the cluster to the language with the higher score.

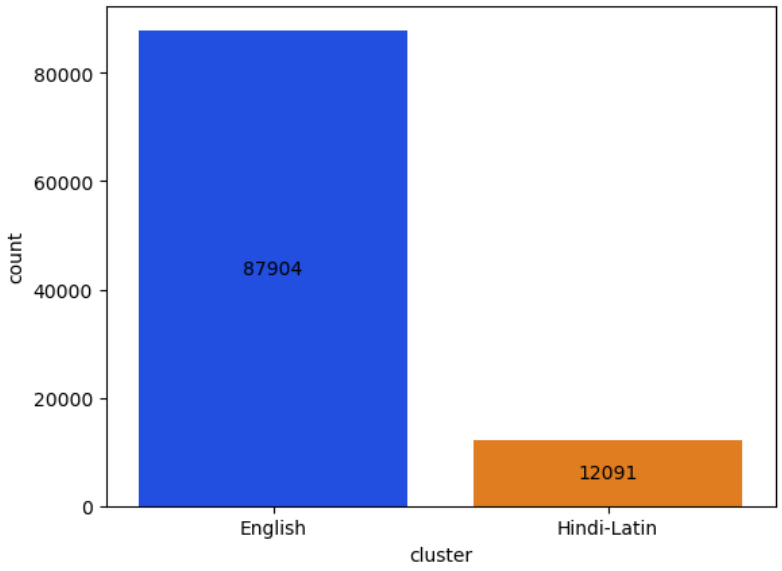


Figure 1.1: Hindi-Latin/English Distribution before cluster reassignment

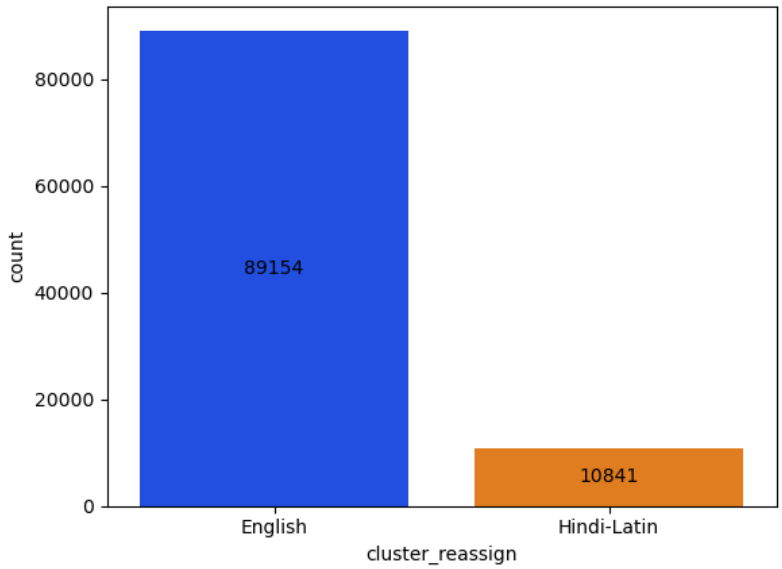


Figure 1.2: Hindi-Latin/English Distribution after cluster reassignment

Before	After	Description
Hindi-Latin	English	Sir i [redacted] of [redacted] i am 12 failed i have to join army i am able in all academic performance written physical and medical but i have no money to repeat 12 standard but my dream is to join army
Hindi-Latin	English	Sir, I am student. I am not terrorist.Ham logo ki maag hai timely exam,timely vacancy,timely results and joining.vacancy ager badhaya n ja raha hai to ghataya bhi n jaye.
Hindi-Latin	English	I have invested [redacted] on [redacted] in Power Bank App which is in play-store, now that app is not available and finally I came to know that it's a scam

Table 1.1: Grievances which were earlier classified as Hindi Latin. Tokens that could reveal personal identifiable information (PII) have been redacted.

Before	After	Description
English	Hindi-Latin	Hello sir Shradheo Naredra modi ji.mujhe upse ek onurodh hey Ki voter counting Jo huya bo firse recounting kijiye.....yea pura [redacted] yehi chahate hey.yea hum un Sab [redacted] lover ka onurodh hey Ki up recounting karne ka order kijiye...
English	Hindi-Latin	Sir Mujhe argent ventilator ki requirement hai oxygen level 50 ho gya hai yahan per koi bi Hospital me bed nahi Hai, ventilator nahin mila to mere dost ke Papa Ki death ho jayegi please sir help kijiye ????????????
English	Hindi-Latin	Sir , Apse anurodh h plz students ki awaz ko sun lijiye .....jobs ko regular kijiye .....jobs ka calender hona chahiye like UPSC.....jldi result aye jisse students confuse n ho .PLEASE HELP US.

Table 1.2: Grievances which were earlier classified as English. Tokens that could reveal personal identifiable information (PII) have been redacted.

1.6 Results

The language detection function was applied to a large number of complaints, resulting in a distribution of the language employed in these complaints. The investigation revealed that Hindi-Latin (Hindi written in the Latin script) was the second most spoken language. Less frequently employed were Punjabi, Gujarati, Telugu, Tamil, Kannada, Odia, and Bengali.

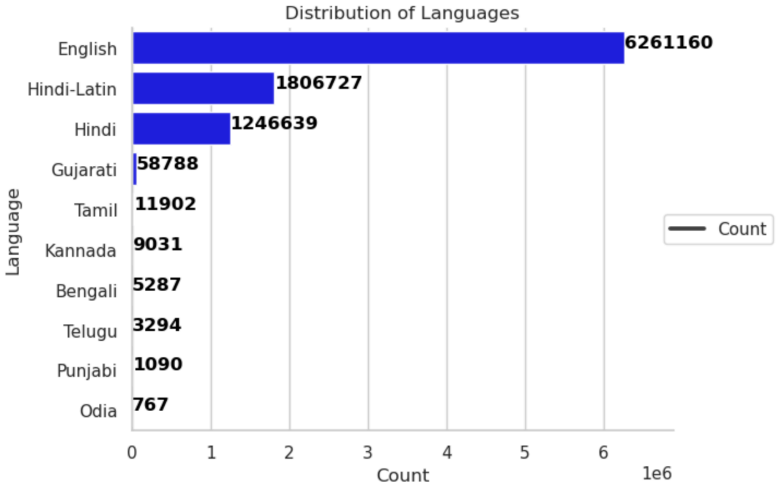


Figure 1.3: Results: Count of various Indian languages

In order to satisfy the diverse linguistic requirements of the population, it is necessary to support multiple languages, which has implications for e-government applications. The language detection function can be used to facilitate cross-lingual semantic queries, allowing users to retrieve relevant complaints in any language.

Despite this, the study had a number of limitations. Initially, only a limited number of languages were considered; future research should investigate the detection of additional languages. No analysis of the content of the complaints was conducted, and future research could examine the relationship between language and the type of complaint being expressed.

In conclusion, the study provided insight into the distribution of languages used in e-governance application-collected complaints and demonstrated the potential for cross-lingual semantic queries. The language detection function can be used to support multilingual e-governance applications; however, additional research is required to resolve the study's limitations and investigate additional languages.

1.7 Conclusion and Future Work

In this chapter, we proposed a language identification method involving the definition of Unicode ranges for different languages, the creation of a language detection function, the detection of the language of each sentence in a dataset, and the clustering of non-Roman script sentences using the KMeans algorithm. We classified English and Hindi-Latin sentences in a dataset using this method and counted the number of documents in each language group. Our method obtained a high degree of accuracy and is capable of

identifying additional languages.

Based on our language identification method, there are multiple avenues for future development. Initially, we can expand the scope of the method by defining Unicode ranges for additional languages. Second, various clustering algorithms can be used to classify non-Roman script sentences into distinct groups. Thirdly, we can investigate the effect of text pre-processing techniques on the accuracy of language recognition. Finally, we can test the scalability and efficacy of our method on broader data sets.

Chapter 2

Examination of Alternate and Conjoined Linguistic Forms

Contents

Abstract *This thesis chapter examines alternate and conjoined linguistic forms. The first section utilizes dynamic programming to find the most probable segmentation of a given input text. The second section creates a dictionary of nearest neighbors for each word in a set of input data using an HNSW index. The third section defines functions for spelling correction using the edit distance algorithm. The fourth section creates a word mapping dictionary to correct words based on the set of letters they contain. The fifth section creates a set of Hindi translations and recursively looks up the correct word using the word mapping. Overall, these sections provide various methods for examining and manipulating linguistic forms.*

2.1 Introduction

The examination and manipulation of linguistic forms play a crucial role in the fields of linguistics and language processing. This chapter of the thesis investigates alternative and combined linguistic forms in an effort to improve the accuracy and efficiency of spell-checking. Dynamic programming is used to determine the most probable segmentation of input text, while an HNSW index is used to generate a dictionary of nearest neighbours for each word in a given dataset. In addition, the edit distance algorithm is used to aid in spelling correction. In addition, a word mapping dictionary is created to correct

words based on their constituent letters, and a recursive method is used to generate Hindi translations and look up the correct word using the word mapping. These sections provide an assortment of methodologies for analysing and manipulating linguistic forms.

This chapter addresses the need for an efficient and accurate spell checker as its central issue. Utilising the power of dynamic programming, the edit distance algorithm, and the HNSW index, the objective is to create a solution that can effectively segment words, generate spelling corrections, and generate a comprehensive dictionary of nearest neighbours for each word. The ultimate goal is to significantly improve the precision and speed of spell-checking processes.

In the following sections, we will delve into the specifics of segmentation, spelling correction, and dictionary creation techniques. The significance of these methods lies in their potential to revolutionise spell-checking mechanisms, thereby enhancing their effectiveness and speeding up the entire process. By investigating alternative and conjoined linguistic forms, we unlock a vast array of opportunities for enhancing language processing applications.

2.2 Problem Statement

Develop an efficient spell checker using dynamic programming, edit distance, and HNSW index for improved accuracy and speed.

2.3 Related Work

1. For the NLP task of segmentation using dynamic programming and Viterbi algorithm, a relevant paper is "Speech and Language Processing" by Jurafsky and Martin (11) (Chapter 5), which provides a comprehensive explanation of the Viterbi algorithm and its application to speech and language processing tasks.

2. For the HNSW index and frequency calculation, a relevant paper is "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs" by Malkov and Yashunin (13) (2018), which introduces the HNSW index and its implementation for efficient nearest neighbor search. The frequency calculation method can be found in various sources, such as "Python Machine Learning" by Raschka and Mirjalili (18) (Chapter 3).

3. For spelling correction using edit distance, a relevant paper is "Fast Approximate String Matching with Large Edit Distances" by Ukkonen (21) (1985), which introduces a method for approximate string matching using a dynamic programming approach with edit distance.

2.4 Methodology

1. For the NLP task of segmentation, use the dynamic programming approach with the Viterbi algorithm to recursively generate all possible splits and calculate the

probability of each split using the Pwords function. (7)

Algorithm 1 Viterbi algorithm for NLP segmentation

procedure VITERBI(x)
  $T(0) \leftarrow$ empty segmentation
  for $i \leftarrow 1$ to n **do**
    $T(i) \leftarrow \bigcup_{j=1}^{i} t \cdot s$ where $t \in T(j-1)$ and s is a segmentation of $x_{j:i}$
    for $y \in T(i)$ **do**
      $P(y|x_{1:i}) \leftarrow 1$
      for $j \leftarrow 1$ to $|y|$ **do**
        $P(y_j|x_{j:i}) \leftarrow$ probability of y_j given $x_{j:i}$
        $P(y_{1:j-1}|x_{1:j-1}) \leftarrow$ maximum probability of a segmentation of $x_{1:j-1}$
        $P(y|x_{1:i}) \leftarrow P(y|x_{1:i}) \cdot P(y_j|x_{j:i}) \cdot P(y_{1:j-1}|x_{1:j-1})$
      end for
    end for
  end for
  $y^* \leftarrow \arg\max_{y \in T(n)} P(y|x_{1:n})$
  segments $\leftarrow$ segments in y^* obtained by backtracking
  return *segments*
end procedure=0

Note that in the above algorithm, we assume that we have access to the probability function $P(y_i|x_{j:i})$ that gives the probability of the ith segment or word in the output given the jth character or word in the input. We also assume that we have a function that can compute the maximum probability of a segmentation of a given sequence. In practice, these functions may be learned from training data using techniques such as maximum likelihood estimation or neural networks.

2. To manipulate a dictionary of Roman words, use the segment function to break down each word into its constituent parts, then multiply the components based on the frequency of the original word using the increase_count function. (22)

3. Train an HNSW index using the fit_hnsw_index function with a set of input features in the "l2" space and initialize it with specified values of "ef" and "M." Use the freq function to calculate the frequency of each alphabet in a given input sentence and return a normalized array of these frequencies. (12)

4. Create a primary dictionary of nearest neighbors for each word in a set of input data by converting histograms stored in a Pandas DataFrame to a list of lists format. Fit an HNSW index to the list of histograms, then use a function to find the nearest neighbors for a given word based on the HNSW index. Call this function for each word in a pre-defined set of words to create a dictionary where each word maps to its nearest neighbors.

5. Use the edit distance algorithm to define functions for spelling correction. (15)

6. Generate a list of possible spelling corrections for a given word, rank them by their probability of appearing in a dictionary, and return the correction with the highest probability using a dictionary of known words.

7. Create a dictionary mapping letters to sets of words containing that letter by reading a list of English words. Remove all sets of words containing the letter 'a'. Create a word mapping dictionary by assigning each word to its corresponding set of letters. Define a function to correct words based on this mapping.

8. Create a set of Hindi transliterations and an empty dictionary to store the final word mapping. These transliterations are created using the method as explained in 2.5.2 subsection. Define a function to recursively look up the correct word using the word mapping. Iterate over the keys in the word mapping, and for each key that is not an English word or a Hindi transliterations, use the get_correct_word function to find the correct word and store it in the final_wording dictionary. If the correct word is already in the dictionary, map the current word to the same value.

2.5 Experiments/Subroutines

2.5.1 Text Processing Pipeline for Finding Words with Minimum Pattern Count: A Generalized Approach

For this experiment, the steps 1-5 of Methodology section 2.4, stay the same. The rest of the process is a under:

1. The pipeline takes as input a dictionary containing words and their nearest neighbors. For each key in the dictionary, the corresponding value (which is a list of words and their nearest neighbors) is retrieved. The list of words is then processed to find the word with the minimum count of a certain character or pattern (in the given code, it is the count of hash symbols '#').

2. To find the word with the minimum count of the chosen pattern, each word in the list is tokenized using a specified tokenizer. The tokenizer breaks each word into tokens based on a specified rule. The resulting tokens are then counted to determine the total count of the chosen pattern in each word. The list of words is then sorted based on the count of the chosen pattern in each word, with the word having the minimum count appearing first.

3. Finally, the list of words is mapped to the word with the minimum count of the chosen pattern. This mapping is done by creating a dictionary with each word in the list as a key and the word with the minimum count of the chosen pattern as its value.

2.5.2 Transliteration of Hindi Text to Roman Script: A Generalized Approach

1. The experiment involves processing text data written in Hindi language.

2. The text data is read from a file with a given file path, and is encoded with 'utf-8' to handle Hindi characters.

3. A counter object is created to keep track of the occurrences of words in the text data.

4. The experiment iterates over each line in the text data, removes leading and trailing whitespaces, and updates the counter object with the word occurrences.

5. Transliteration is performed on a specific Hindi word "नमस्ते" (meaning "Hello" in Hindi) using a transliteration function with specified language code and transliteration options.

6. The transliterated word is stored in a dictionary with the original word as the key and the transliterated word as the value.

7. Further, the experiment iterates over all the keys (words) in the counter object, transliterates each word using the same transliteration function with the same language code and options, and stores the transliterated words in the dictionary.

2.5.3 Comparing Tokenizers for Segmenting Hindi Latin Words and Sentences Based on Entropy Analysis

This code is designed to evaluate the entropy of different tokenizers when applied to a dataset of Hindi grievances. The four tokenizers used in this evaluation are

1. salesken/similarity-eng-hin_latin (4)

2. l3cube-pune/hing-bert (2)

3. l3cube-pune/hing -mbert (3)

4. facebook/m2m100_418M (1)

We initialize these tokenizers using the AutoTokenizer.from_pretrained() and M2M100Tokenizer.from_pretrained() methods, and then tokenize the dataset of Hindi grievances using these tokenizers. We define two functions, count_chars(word) and count_freq(s, tokeniser), to tokenize sentences and count the frequency of lowercase English alphabets in each token. We also define the entropy(labels, base=None) function to calculate the entropy of a list of labels.

For each sentence in the dataset, we apply each tokenizer using a for loop and calculate the entropy of the list of counts of lowercase English alphabets in each token using the entropy() function. We then calculate the average entropy for each tokenizer by summing up the entropy values of all sentences and dividing by the total number of sentences.

A dictionary called out is created to store the average entropy values for each tokenizer, and the keys of the dictionary are the names of the tokenizers. Finally, we sort the dictionary out based on the entropy values in ascending order using a lambda function, and return the sorted dictionary as the output of the process.

Based on the results obtained from the comparison of four tokenizers (salesken/similarity-eng-hin_latin, l3cube-pune/hing-bert, l3cube-pune/hing-mbert, and facebook/m2m100_418M)

Tokenizer	**Average Entropy over 500 Hindi-Latin Documents**
salesken/similarity-eng-hin_latin	2.2282
l3cube-pune/hing-bert	1.9941
l3cube-pune/hing-mbert	2.0253
facebook/m2m100_418M	1.9701

Table 2.1: Comparison of Tokenizers based on Average Entropy over 500 Hindi-Latin Documents

using average entropy over 500 Hindi-Latin documents, it can be concluded that the facebook/m2m100_418M tokenizer has the lowest average entropy of 1.9701, while the salesken/similarity-eng-hin_latin tokenizer has the highest average entropy of 2.2282. The l3cube-pune/hing-bert and l3cube-pune/hing-mbert tokenizers have average entropies of 1.9941 and 2.0253, respectively.

Therefore, if the goal is to perform segmentation of Hindi-Latin words and sentences, the facebook/m2m100_418M tokenizer may be the most suitable option, based on its lower average entropy. However, the choice of tokenizer ultimately depends on the specific use case and the performance of the tokenizer on the particular data being analyzed.

2.6 Results

```
systeym: system          bethtar: behtar          verbually: verbally
systsem: system          gambbhir: gambhir        verballby: verbally
sysytem: system          gambhri: gambhir         vervally: verbally
sysetem: system          gambhira: gambhir        verbvally: verbally
behatr: behtar           gammbhir: gambhir        verdbally: verbally
bethar: behtar           gambhier: gambhir        bhaaut: bahut
behetar: behtar          sabhapatti: sabhapati    bhahaut: bahut
behttar: behtar          sabhapati: sabhapati     bbahaut: bahut
behthar: behtar          sabhapatis: sabhapati    bahatu: bahut
bhetar: behtar           sabhapat: sabhapati
```

Listing 2.1: Sample of Final Word Mappings

2.7 Conclusion and Future Work

In this chapter of the thesis, we discussed a variety of natural language processing (NLP) techniques and algorithms for solving a variety of problems, including segmentation, dictionary manipulation, nearest neighbour search, spelling correction, and word mapping for multiple languages. We discussed dynamic programming, the Viterbi algorithm, the HNSW index, the edit distance algorithm, and a number of other techniques for completing these tasks. In addition, we discussed the implementation specifics and steps for each method.

Although we have covered a wide variety of NLP techniques and algorithms, there is still room for additional research and development in this field. For instance, the use of deep learning models such as transformers for various NLP tasks can be investigated. In addition, there is still much work to be done to improve the accuracy of existing algorithms and techniques. Extending the techniques covered in this chapter to other languages and scripts is an additional intriguing direction for future research.

This chapter concludes with a comprehensive overview of diverse NLP techniques and algorithms for various tasks. It lays the groundwork for future research and development in this area, and we hope it will inspire researchers to explore novel and inventive approaches to NLP problem-solving.

Chapter 3

Efficient Text Clustering for Multi-Lingual Bulk Campaigns

Contents

Abstract *This chapter addresses the issue of bulk campaigns in large datasets of grievances. To solve this problem, we propose a method that clusters similar grievances using text embeddings and the Hierarchical Navigable Small World (HNSW) index. The method identifies the representative document for each bulk and returns a mapping containing the representative document ID and the length of the bulk for each document in the input data.*

3.1 Introduction

Text document clustering is a challenging problem, especially when working with large datasets containing multiple languages. Particularly challenging are bulk campaigns, which can generate significant noise and make it difficult to effectively query a database. In this chapter, we propose a solution to this problem by introducing a novel function that groups similar documents according to their contents. The function generates clusters of similar documents using word embeddings and a Hierarchical Navigable Small World (HNSW) index from a Pandas DataFrame containing text documents. This method enables an efficient k-nearest neighbour search, allowing us to quickly identify documents with similar content. Additionally, the function is ideal for use in multilingual environments because it can process text documents written in multiple languages. By clustering a large dataset of complaints, we demonstrate the efficacy of our method and that our function can efficiently and precisely identify clusters of related documents.

3.2 Problem Statement

Identify bulk campaigns of similar documents within a large corpus of multilingual text data

3.3 Related Work

Several research papers have explored the use of multilingual word embeddings and efficient indexing methods for similarity search.

Conneau et al. (9) proposed a method for training multilingual word embeddings using aligned multilingual corpora from Wikipedia texts. They achieved state-of-the-art results on various cross-lingual evaluation tasks. This chapter also uses pre-trained word embeddings from Wikipedia texts for multiple languages, but instead of training the embeddings from scratch, it uses pre-trained embeddings and creates a unified index using a Hierarchical Navigable Small World (HNSW) index.

Malkov et al. (14) introduced the HNSW graph, a method for building a graph structure that enables fast approximate nearest neighbor search. The HNSW graph is used in this chapter to create a unified index of pre-trained word embeddings for multiple languages, which enables efficient similarity search across languages.

Mikolov et al. (16) proposed the Word2Vec algorithm for learning distributed representations of words in a vector space. The authors presented two methods for training word embeddings: continuous bag-of-words (CBOW) and skip-gram. This chapter uses pre-trained word embeddings generated using the Word2Vec algorithm and creates a unified index using an HNSW graph.

Artetxe et al. (6) presented an empirical comparison of several methods for training cross-lingual word embeddings, including using parallel corpora, using cross-lingual signal propagation, and using a combination of monolingual and cross-lingual information. The authors evaluated the embeddings on a range of cross-lingual tasks and found that the combination of monolingual and cross-lingual information performs best. The approach presented in this chapter uses pre-trained word embeddings generated from Wikipedia texts for multiple languages, which can be seen as a form of cross-lingual information integration.

3.4 Methodology

1. Apply a text cleaning function called *clean_text* to the ”subject_content” column of the input DataFrame to remove any unwanted characters or noise.

2. Generate word embeddings for each text using the *generate_embedding* function, which takes a sentence and a pre-trained embeddings dictionary as input and returns a 300-dimensional vector representation of the sentence.

3. The method to build embeddings dictionary loads pre-trained word embeddings for nine different languages (English, Tamil, Telugu, Kannada, Hindi, Bengali, Gujarati,

Punjabi, and Oriya) using the gensim library in Python. It creates an empty dictionary called embeddings_dict, and then loops through all the files in a specified folder path to check if each file is a .vec file. If it is, it loads the embeddings from the file using gensim and adds them to the embeddings_dict dictionary using the file name as the key. This dictionary will be used later in the clustering process to generate embeddings for the text data and identify similar documents. By leveraging pre-trained word embeddings for multiple languages, this chapter's approach is able to efficiently process multilingual text data.

4. Use the embeddings to fit a Hierarchical Navigable Small World (HNSW) index with an exploration factor (ef) of K*10, where K is set to the minimum of the length of the DataFrame and 1000. The HNSW index allows for efficient k-nearest neighbor search, which is used to identify similar documents based on their embeddings.

5. Find the K nearest neighbors and their distances for each embedding, and calculate the L2 distance threshold based on a cosine similarity threshold (which is not used in this implementation). Create a dictionary where each key is an index of an embedding and its value is a list of indices of its nearest neighbors within the L2 distance threshold. This dictionary is used to identify clusters of similar documents.

6. Create an empty list and set to store all the bulk sets and indices that have been added to a bulk. Loop through the primary documents in the primary_dict and for each primary document, create a temporary list called *tmp_docs* that contains the primary document and all of its similar documents within the L2 distance threshold. If the *tmp_docs* list contains more than *min_docs* documents, add it to the *bulk_all* list.

7. Initialize a dictionary to store the bulk status for each document, loop through each bulk in the *bulk_all* list, sort the documents in the bulk in ascending order, and assign the first document in the sorted bulk as the "representative" document for all documents in the bulk. Set the output for each document to a tuple of the representative document ID and the length of the bulk.

8. Return the output dictionary containing the representative document ID and the length of the bulk for each document in the input DataFrame.

3.5 Results

3.5.1 L2 Distance and EF

Exploration factor (ef): The exploration factor determines the number of nodes to be visited during the graph traversal process. It is a parameter that affects the trade-off between search speed and search quality. A higher ef value results in a more accurate search at the cost of longer search times. The ef value can be calculated as $K * 10$, where K is the minimum of the length of the DataFrame and 1000. This calculation ensures that the ef value scales with the size of the data and provides a reasonable balance between search quality and search efficiency.

L2 distance threshold: The L2 distance threshold is the maximum distance allowed between two points in the vector space for them to be considered similar. It is calculated based on the cosine similarity threshold, which is a threshold used to determine whether two vectors are similar or dissimilar. To calculate the L2 distance threshold, first, the cosine similarity threshold is converted to an L2 distance threshold using the following formula:

$$l2_{threshold} = \sqrt{2 * (1 - cosine_{threshold})}$$

Once the L2 distance threshold is calculated, the K nearest neighbors and their distances for each embedding are found using the HNSW index. The L2 distance threshold is used to filter out any nearest neighbors that are further away than the threshold. A dictionary is then created where each key is an index of an embedding, and its value is a list of indices of its nearest neighbors within the L2 distance threshold. This dictionary is used to identify clusters of similar documents.

3.5.2 Example of a Bulk Campaign

3.5.2.1 Sampled List of Grievances

reg_no	subject_content
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted] Name of Bank and Branch : [redacted] of Application : [redacted]----------To, The [redacted] [redacted]ବିଷୟ: ଜଣଟୋମାଳ ବ୍ଲକର ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସର ଜୌଣସି ଜାମ କରୁ ନେ ଥିବାରୁ ଫିଲ୍ଡ ଅଫିସରଙ୍କୁ ଖୁବ୍ ଶୀଘ୍ର ବଦଳି କରାଯାଇ ଜଣ୍ଟୋମାଳ ବ୍ଲକର ଗରୀବ ଯୁବକ ମାନଙ୍କ ଜରିଆରେ ଫିଲ୍ଡ ଅଫିସର ବଦଳୀ ପଲାଖେ।ମାନାଶୟ, ଲେଖି ଜଣାଇବାର ଜାରଣ ଏହି କି ଯେ ବିଗୟତ ଖନ ଜଣ୍ଟୋମାଳ ବ୍ଲକର ଗରୀବ ଯୁବକ ମାନେ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବାରମ୍ବାର ଫିଲ୍ଡ ଅଫିସର ରିଯେଷ୍ଟ କରିଥିବାରୁ ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ [redacted] ଫିଲ୍ଡ ଅଫିସର ବଦଳି କରା ଯାଇ ନାହିଁ। ଫିଲ୍ଡ ଅଫିସର ଉଙ୍କ ପରୁ ଧ୍ୟାନ ବିକ୍ରୀ କରି ସରକାରୀ PMEGP ଋଣ ପ୍ରଦାନ କରୁଛନ୍ତି ଯେ ବାରମ୍ବାର ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବିଭିନ୍ନ ପ୍ରକାର ନାଟ ଦେଖାଇ ଗରୀବ ଯୁବକ ମାନଙ୍କ CIBIL score low କରି ବାରମ୍ବାର ରିଯେଷ୍ଟ କରିଥିବାରୁ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଜାଗାର ପତ୍ତର ଦାଖଲରେ ଖର୍ଚ୍ଚ ହୋଇ ଥିବା ତଙ୍କା ବ୍ୟାପକ ଭାବେ ମାନାନି କ୍ଷତି କରିଥିବା ଫିଲ୍ଡ ଅଫିସର [redacted] [redacted] ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ ତୁରନ୍ତ ବଦଳି ଦାବିରେ।ଅନ୍ୟ ଲୋକଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ CIBIL score ବଢ଼ାଇ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି PMEGP ଋଣ ପ୍ରଦାନ କରୁଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କ CIBIL ସ୍କୋରରେ ନାଟ କରି ସମସ୍ତ ପ୍ରକାର ଜାଗାର ପତ୍ତର ବର୍ଷ ବର୍ଷ ଧରି pending ରଖି ଆଜି ଦେଖିବା ଜାଲି ଦେଖିବା ଜଣି ବାରମ୍ବାର ରିଜେଷ୍ଟ କରୁଥିବା ଫିଲ୍ଡ ଅଫିସର ବୈଦ୍ଧ ଜିଲ୍ଲାର [redacted] [redacted] ରହିବାର ଆବଶ୍ୟକତା ତୁରନ୍ତ ବଦଳି କରାଯାଉ ବୋଲି ଜଣ୍ଟୋମାଳ ବ୍ଲକର ଗରୀବ ଯୁବକ ମାନେ ଦାବି ପତ୍ତର ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ସରକାରୀ ଅଫିସର୍ ଋଣ ଜାଡିକି ବଦଳି କରୁନାହିଁନ୍ତି ଜବାବ୍ ଦିଅନ୍ତୁ ।PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରୁ ଜିଛି ଅଫିସର୍ ମାନେ ବ୍ୟାପାର ଜରି ମଦ ମାଂସ ଖାଇ ଜମିଶାନ ନେଉ ଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନେ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖାଲ କରୁଥିଲେ ମଧ୍ୟ ସମସ୍ତ ତଥ୍ୟ ଗୁଡ଼ିକ ପ୍ରଦାନ କରିବା ପାଇଁ ସବୁ ତଥ୍ୟ ଗପି ରଖିଛନ୍ତି ।PMEGP ଋଣ ଯୋଜନା ଉପରେ [redacted] [redacted] ଚୁଦ୍ ଚଣ୍ଡାଳ୍ ମାନଙ୍କୁ PMEGP [redacted] କରୁଥିବା ବେଳେ ସମସ୍ତ ପ୍ରକାର ଲିଖିତର ତଥ୍ୟ PMEGP ପୋର୍ଟାଲ ମଧ୍ୟମାରୁ ପ୍ରିଣ୍ଟ କରି ପ୍ରଦାନ କରିପାରୁ ନଥିବା ବେଳେ ବ୍ୟାପକ ଭାବେ ଲୋଟି ଲୋଟି ତଙ୍କା ଦୁର୍ନୀତି କରି ସମସ୍ତ ତଥ୍ୟ ଗପା ରଖୁଛନ୍ତୀ [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସର। ଏଣୁ ଆପଣଙ୍କୁ ବିନୀତ ଅନୁରୋଧ କରୁଛି ଯେ ଉପରୋକ୍ତ ବିଷୟ ପ୍ରତି ଧ୍ୟାନ ଦେଇ ଖୁବ୍ ଶୀଘ୍ର ପଦକ୍ଷେପ ନିଆଯାଉ ।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted] / Name of Bank and Branch : [redacted]Date of Application : [redacted]----------To, The [redacted] [redacted]ସରକାର ଜାଡିକି ଗରିବ ଯୁବକ ଓ ଯୁବତୀ, ଜଣ ସାଧାରଣ ଲୋକ ମାନଙ୍କ ପାଇଁ PMEGP ଋଣ ଯୋଜନା ପ୍ରବର୍ତ୍ତନ କରି ଛନ୍ତି, ସରକାରୀ ଯୋଜନାରେ ସାମିଲ କରି ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ମ ବର୍ଷ ବର୍ଷ ଧରି ଆଣ୍ଠା ବେର ସମସ୍ତ ପ୍ରକାର ଜାଗାର ପତ୍ତର ପକାର ରଖି ଥୋର ଋଣ କରିବା ପାଇଁ ସରକାର PMEGP ଯୋଜନା ପ୍ରବର୍ତ୍ତନ କରିଛନ୍ତି ।ଏ ବିଷୟରେ ବାବୁ ଅଫିସର ମାନେ ଟିକେ ଉତ୍ତର ଦେବେକି।ସରକାର ଫିଲ୍ଡ ଅଫିସର ଏବଂ ବ୍ୟାଙ୍କ ମ୍ୟାନେଜରଙ୍କୁ ଜାହିଁକି ନିର୍ଦ୍ଦେଶ ଦେଇ ଛନ୍ତି, ଗରିବ ଯୁବକ ଓ ଯୁବତୀ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ଋଣ ଆବେଦନ ଜଲେ ତମକୁ ଅଫିସର ମାନଙ୍କୁ ଜାହିଁ ଭିଡ଼ ପଡ଼ୁଛେ ବ ତୁମର ଉରମା ଚଟ୍ଜାରୀ PMEGP ଋଣ ପ୍ରଦାନ କରୁଛି କି ଯେ ଏତେ ଭିଡ଼ ପଡୁଛେ।ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାର ସାହାଯ୍ୟ କରୁଥିଲେ ମଧ୍ୟ ଜଚଜିଆ ଅଫିସର ମାନେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ଜୁଟିଗଲି ଛନ୍ତି [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ସରକାରି PMEGP ଋଣ ପ୍ରଦାନ ହେବା ପାଇଁ, ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ବୈଦ୍ଧ ଜିଲ୍ଲାରେ ସରକାରି ଯୋଜନାରେ ଲୋଟି ଲୋଟି ତଙ୍କା ଜୁଟି ଅନ୍ୟ ଲୋଟି ପୋଟି ଲୋକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି ମଦ ମାଂସ ଭୋଗି ଖାଇ ଜାମ କରି ଗରୁ ଥିବା ବେଳେ ଗରିବ ଯୁବକ ମାନଙ୍କୁ ବହୁତ୍ ଅନ୍ୟାୟ କରି ଗରିଥିବା ଅଫିସର ମାନଙ୍କୁ ଡିସେମ୍ବର ମାସ ସୁଦ୍ଧା ବଦଳି ବରଜାରୁ।ଲୋଟି ଲୋଟି ତଙ୍କା ଜୁଟୁ ଥିବାରୁ PMEGP ପୋର୍ଟାଲ ସାଇଟ୍ କୁ PMEGP ଋଣ ଆବେଦନ କରିଥିବା ସମସ୍ତ ତଥ୍ୟ ଦେଇ ପାରୁ ନାହିଁ, ବାରମ୍ବାର ମିଥ୍ୟା ତଥ୍ୟ ଦେଇ ଗଲି ଛନ୍ତି ତୁରନ୍ତ ସମସ୍ତ ଲିଷ୍ଟର ଆବଶ୍ୟକ [redacted] [redacted] [redacted] ମସିହା ପର୍ଯ୍ୟନ୍ତ ଋଣ ନେଇ ଥିବା ଆବେଦନ ଜାଗାଙ୍କ ତଥ୍ୟ ବରଜାର୍।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted] / Name of Bank and Branch : [redacted] Date of Application : [redacted]----------To, The [redacted] finance minister department government of India [redacted]ବିଷୟ: [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସରଙ୍କୁ ନିର୍ଦ୍ଦେଶ ଦିଅନ୍ତୁ ଯେ ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ରିଯେଷ୍ଟ କରିଥିବା ପୁନର୍ବାର [redacted] [redacted] re forward କରି ଗରୀବ ଯୁବକ ମାନଙ୍କ ଋଣ ପ୍ରଦାନ କରା ଯାର ନେ ହେଲେ ଆଗକୁ ଆହୁରି ଅନେକ ଅଭିଯୋଗ ଦାଖଲ ଜାରୀ ରହିବ ଫିଲ୍ଡ ଅଫିସରଙ୍କ ବିରୁଦ୍ଧରେ ଦୃଢ଼ ଜାର୍ଯ୍ୟାନୁଷ୍ଠାନ ଗରହଣ କରାଯାଉ । ଫିଲ୍ଡ ଅଫିସର ଯଦି [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ସବୁ ଦିନ ଜାମ କରିବା ପାଇଁ ଗଡ଼ୁଛି ତାଲେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଋଣ ପ୍ରଦାନ କରା ଯାର ନେ ହେଲେ ଅଭିଯୋଗ ସବୁ ଦିନ ପାଇଁ ଜାରି ରହିବ।ମାନାଶୟ, ଲେଖି ଜଣାଇବାର ଜାରଣ ଏହି କି ଯେ [redacted] ପ୍ରଥ୍ୟା ବାସିନ୍ଦା ଖୋଇ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ପାଇଁ ଋଣ ପ୍ରଦାନ ଖୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ନାଟ କରି ରାଜନୀତି କରି ଜାଗାଜ ଜଲ୍ପାରେ ଅଟକାଇ ରଖିଥିବା ଏବଂ ରିଯେଷ୍ଟ କରିଥିବା ଋଣ ଫିଲ୍ଡ ଅଫିସର ପୁନର୍ବାର [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କକୁ re forward କରି ଋଣ ପ୍ରଦାନ କରା ଯାଉ ନେ ହେଲେ ଆଗକୁ ଆହୁରି ଅନେକ ଅଭିଯୋଗ ଦାଖଲ ଜାରୀ ରହିବ ଫିଲ୍ଡ ଅଫିସରଙ୍କ ବିରୁଦ୍ଧରେ, ବାରମ୍ବାର ଅଭିଯୋଗ କରୁଥିଲେ ମଧ୍ୟ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ଋଣ ପ୍ରଦାନ ଖୋଇ ପାରୁନାହିଁ ତୁରୁ ମାନଙ୍କୁ ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ଋଣ ପ୍ରଦାନ କରୁଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ବର୍ଷ ବର୍ଷ ଧରି ନାଟ ଦେଖାଇ ଆଜି ମଞ୍ଜୁର ହେବା ଜାଲି ମଞ୍ଜୁର ହେବା ଜଣି ବ୍ୟାପକ ଭାବେ ମାନାନି କ୍ଷତି କରିଥିବାରୁ ମାନାନି କ୍ଷତି ଟଙ୍କା ଭରଣ ପୋଷଣ ନେ କରିଛ ସେ ପର୍ଯ୍ୟନ୍ତ ଅଭିଯୋଗ ଜାରି [redacted] ଜଲ୍ପମାନ [redacted]ମାଓ ବାଦି ମାନଙ୍କ ଜରିଆରେ ଚେତବନୀ ତୁରନ୍ତ ଫିଲ୍ଡ ଅଫିସର ବଦଳି ଦାବିରେ ବୈଦ୍ଧ ଜିଲ୍ଲାର ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଫିଲ୍ଡ ଅଫିସର ଜୌଣସି ଜାମ କରୁ ନେ ଥିବାରୁ ବଦଳି ଦାବିରେ।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted] / Name of Bank and Branch : [redacted] Date of Application : [redacted]----------To, The [redacted] [redacted]ବିଷୟ: [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ଥିବା ଫିଲ୍ଡ ଅଫିସର ସରକାରୀ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବଡ଼ ବଡ଼ିଆ ଲୋକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି ଋଣ ପ୍ରଦାନ କରୁଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ବର୍ଷ ବର୍ଷ ଧରି ଆଶ୍ୱସନା ଦେଇ ଆଜି ଦେଖିବା ଜାଲି ଦେଖିବା ଜଣି ସମସ୍ତ ପ୍ରକାରର ଜାଗାଜ ପତ୍ତର ବର୍ଷ ବର୍ଷ ଧରି [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ପକାଇ ରଖୁଥିବା ଫିଲ୍ଡ ଅଫିସର [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ, ଅନ୍ୟ ଲୋକଙ୍କ ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରୁ ଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଜାମ ନେ କରୁଥିବା ଫିଲ୍ଡ ଅଫିସର ବୈଦ୍ଧ ଜିଲ୍ଲାରୁ ତୁରନ୍ତ ବଦଳି କରାଯାଉ ।CIBIL ସ୍କୋରରେ ବିଭିନ୍ନ ପ୍ରକାରର ନାଟ ଦେଖାଇ ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବାରମ୍ବାର ରିଯେଷ୍ଟ କରୁଥିବା ବେଳେ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ଋଣ ପ୍ରଦାନ ଖୋଇ ପାରୁନାହିଁ, ସରକାର PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବ୍ୟାଙ୍କଙ୍କ ଗୁଡ଼ିକୁ ନିର୍ଦ୍ଦେଶ ଦେଇଥିଲେ ମଧ୍ୟ ସରକାରଙ୍କ କଥାକୁ ଫ୍ୟୁ କରୁଥିବା ଅଫିସର ବୈଦ୍ଧ ଜିଲ୍ଲାରେ ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ।ମାନାଶୟ, ଲେଖି ଜଣାଇବାର ଜାରଣ ଏହି କି ଯେ [redacted] ସ୍ୱାଥ୍ୟା ବାସିନ୍ଦା ଖୋଇ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ସ୍ୱାବଲମ୍ବୀ ହେବା ପାଇଁ ଜାହିଁକି PMEGP ଋଣ ପ୍ରଦାନ ଖୋଇ ପାରିବ ନାହିଁ, ସରକାର PMEGP ଋଣ ଯୋଜନା ଜାହିଁକି ପ୍ରବର୍ତ୍ତନ କରିଛନ୍ତି, ଋଣ ନେଇ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ମ ବର୍ଷ ବର୍ଷ ଧରି ଆଶ୍ୱସନା ଦେଇ ସମସ୍ତ ପ୍ରକାର ଜାଗାଜ ପତ୍ତର ପେଣ୍ଡିଙ୍ଗ୍ ରଖି ଖୋଇଗଲା ଜରିବା ପାଇଁ ସରକାର ଯୋଜନା ପ୍ରବର୍ତ୍ତନ କରିଛନ୍ତି କି ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ବୈଦ୍ଧ ଜିଲ୍ଲାରେ ଅଫିସର୍ ମାନେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଉପରେ ଅନ୍ୟାୟ କରୁଥିବା ବେଳେ ବାରମ୍ବାର ଅଭିଯୋଗ ଜରେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ମିଳି ପାରୁନାହିଁ, ଆଉ ଜେତେ ବର୍ଷ ଅଭିଯୋଗ ଜଲେ ନ୍ୟାୟ ମିଳି ପାରିବ ଏ ବିଷୟରେ ଉତ୍ତର ପ୍ରଦାନ କରାଯିବ କି, ଗରୀବ ଯୁବକ ମାନେ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ବାରମ୍ବାର ଅଭିଯୋଗ ଜଲେ ବୁଢି ପାରୁ ନାହିଁ ଆଉ ଜେତେ ବର୍ଷ ଅଭିଯୋଗ ପରେ ବୁଢୀ ପାରିବେ ସାର୍।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted] Branch / Name of Bank and Branch : [redacted]Date of Application : [redacted]----------To, The [redacted] [redacted] [redacted] ବ୍ୟାଙ୍କରେ ସରକାରୀ PMEGP ଯୋଜନା ଉପରେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଉପରେ ବ୍ୟାପକ ଭାବେ ଅନ୍ୟାୟ କରୁଥିବା ବେଳେ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ପ୍ରଦାନ ଖୋଇ ପାରୁନାହିଁ, ଆଉ ଜେତେ ବର୍ଷ ଅଭିଯୋଗ ଜଲେ ଗରୀବ ଯୁବକ ମାନେ ନ୍ୟାୟରୁ ବଞ୍ଚି ପାରିବେ , [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ଋଣ [redacted] ପାଇଁ ବାରମ୍ବାର ରିଯେଷ୍ଟ କରିଥିବା ଫିଲ୍ଡ ଅଫିସର ବୈଦ୍ଧ ଜିଲ୍ଲାରେ ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ ତୁରନ୍ତ ବଦଳି ଦାବି।ମାନାଶୟ, ଲେଖି ଜଣାଇବାର ଜାରଣ ଏହି କି ଯେ ବୈଦ୍ଧ ଜିଲ୍ଲାର ସ୍ୱାଥ୍ୟା [redacted] ଖୋଇ ଗରୀବ ଯୁବକ ମାନେ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ଋଣ ପ୍ରଦାନ ଖୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ନାଟ କରି ଜାଗାଜ ଜଲ୍ପାରେ ଅଟକାଇ ରଖିଥିବା ଋଣ ଖୁବ୍ ଶୀଘ୍ର ପ୍ରଦାନ କରାଯାଉ ନେ ହେଲେ ଅଭିଯୋଗ ସବୁ ଦିନ ପାଇଁ ଜାରୀ ରହିବ।ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ଅଫିସର ମାନେ ବୈଦ୍ଧ ଜିଲ୍ଲାର ଗରୀବ ଯୁବକ ମାନଙ୍କ ଉପରେ ବହୁତ୍ ଅନ୍ୟାୟ କରୁଥିବାରୁ ବାରମ୍ବାର ଅଭିଯୋଗ ଜଲେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ପ୍ରଦାନ ଖୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ବିଭିନ୍ନ ପ୍ରକାର ନାଟ କରି ରାଜିନ୍ତି [redacted] [redacted] [redacted] ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସର ।ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ବୈଦ୍ଧ [redacted] ଜଣ୍ଟୋମାଳ ବ୍ଲକର [redacted] ଷ୍ଟେଟ ବ୍ୟାଙ୍କରେ ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ଯୋଜନା ଉପରେ ବହୁତ୍ ଅନ୍ୟାୟ କରୁଥିବାରୁ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ପ୍ରଦାନ ମିଳି ପାରୁନାହିଁ ତୁରନ୍ତ ସମାଧାନ ଦାବି ନେ ହେଲେ ଅଭିଯୋଗ ସବୁ

Figure 3.1: Sampled List of Grievances - 1. Tokens that could reveal personal identifiable information (PII) have been redacted.

	ଦିନ ପାଇଁ ଜାଗା ରହିବ।ଗୋର ଅଫିସରଙ୍କୁ ଉରେଇ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଜଦ୍ୱନାଗଣ୍ଡି ବୈଦ୍ୟ ଜିଲ୍ଲାର ଜଣ୍ଟାମାଳ ବ୍ଲକର ଗରାଁବ ଯୁବକ ମାନେ, ଫିଲ୍ଡ ଅଫିସର ଗରାଁବ ଯୁବକ ମାନଙ୍କ ବ୍ୟାପକ ଭାବେ ମାନହାନି କ୍ଷତି କରି ବାରମ୍ବାର PMEGP ରିଜେକ୍ଟ କରିଥିବାରୁ ଗରାଁବ ଯୁବକ ମାନଙ୍କ ଅଧିକାର ଛଡେଇ ନେଉ ଥିବା ଫିଲ୍ଡ ଅଫିସର ବଦ୍ଧତ୍ ଅନ୍ୟାୟ କରୁଛନ୍ତି ।ଗରାଁବ ଯୁବକ ମାନଙ୍କୁ ଜେଲ୍ ଫାଣ୍ଡା ହେଇ ଯାଉ ପଛେ ଗରାଁବ ଯୁବକ ମାନଙ୍କ ଜୀବନ ପଛେ ଗଲିଯିବ ସରକାରୀ PMEGP ଯୋଜନା ଉପରେ ଲଢ଼େଇ ଜାରି ରହିବ ଯେ ପର୍ଯ୍ୟନ୍ତ ଗରାଁବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ନେ ମିଳିଛି ସେ ପର୍ଯ୍ୟନ୍ତ ଅଭିଯୋଗ ଜାରି ରହିବ।ଦେଖା ଯାଉ ଜେଡ଼େ ଜେଡ଼େ ବଡ ଜାତୁ ଆ ଫିଲ୍ଡ ଅଫିସର ମାନେ [redacted] ଜିଲ୍ଲାରୁ ଆସି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କକୁ ଗଜିରି କରିବାକୁ ଆସୁଛନ୍ତି ଦେଖା ଯାଉ ଅଭିଯୋଗ ଜାରି ରହିବ।ଜଣ୍ଟାମାଳ ବ୍ଲକର ଗରାଁବ ଯୁବକ ମାନଙ୍କ ଜରିଆରେ ଚେତାବନୀ।
DEABD/E/2023/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted]Branch / Name of Bank and Branch : [redacted]Date of Application : [redacted]----------To, The [redacted] ସମ୍ମାନୀୟ ସାର୍ ମୁଁ ଜାନୁୟାରୀ [redacted] ରେ ସମସ୍ତ ପ୍ରକାର ଜାଗାର ପତ୍ର [redacted] ବ୍ୟାଙ୍କ ରେ ଜମା କରି ଅଛି ହେଲେ ବର୍ଷେ ହେବ ବିତି ଯାଇଥିଲେ ସୁଦ୍ଧା ଜୌଣସି ଉତ୍ତର ମିଳି ନାହିଁ, ଆଉ ଜେତେ ବର୍ଷ ଅଭିଯୋଗ ଜଲେ ମୋର୍ ଋଣ ମଞ୍ଜୁର ହୋଇ ପାରିବ ସାର୍, ମୁଁ ବ୍ୟାଙ୍କ କୁ ଯାଇ ଫିଲ୍ଡ ଅଫିସରଙ୍କୁ ପଗରିବାରୁ [redacted] ଅଫିସର କହିଲେ ମୁଁ ବୈଦ୍ୟ ଖଟେଟ୍ ବ୍ୟାଙ୍କ୍ Racc କୁ ସମସ୍ତ ପ୍ରକାର ଜାଗାର ପତ୍ର ଏବଂ ଋଣ ପ୍ରସୋଗ୍ରସ୍ ପଠାଇ ଦେଇଛି ବୋଲି ଫିଲ୍ଡ ଅଫିସର ବର୍ଷେ [redacted] ଠରୁ ଥିବାରୁ ଜୌଣସି ଉତ୍ତର ଆସି [redacted] ପାଖରୁ, ମୁଁ [redacted] ପ୍ରତିଷ୍ଠାନ ପାଇଁ ବାରମ୍ବାର ସମସ୍ତ ପ୍ରକାର ଜାଗାର ପତ୍ର ଦାଖଲ ଜଲ୍ ପରେ ବି ସୁଦ୍ଧା ଜୌଣସି ସୁଫଳ ମିଳି ନଥିବାରୁ ମୁଁ ଜାଗାର ପତ୍ର ରେ ପଗଣ ଜାଗର ଟଙ୍କା ଖର୍ଚ୍ଚ କରି ଅଛି ସେ ଟଙ୍କା ବେବ କିଏ, ମୋର ବଦ୍ଧତ କ୍ଷତି କରିଥିବା ଫିଲ୍ଡ ଅଫିସର ମୋତ୍ କ୍ଷତି ପୂରଣ ଦିୟା ଗଲେ [redacted] ସବୁ ଦିନ ପାଇଁ ବନ୍ଧ ରହିବ ନେ ହେଲେ ଅଭିଯୋଗ ସବୁଦିନ ପାଇଁ ଜାରି ରହିବାର ସୁନ୍ନା ଭୋଇ ତେତ ବନି ଦେଉଛି । ଗୋଟିଏ ଛଟ ସାଖ୍ୟୟ୍ୟ ପାଇଁ ବର୍ଷ ବର୍ଷ [redacted] ହେଲେ ଆଉ ଜେତେ ବର୍ଷ ପରେ ମୋର୍ ଋଣ ମଞ୍ଜୁର ହୋଇ ପାରିବ ଉତ୍ତର ଦେଇ [redacted] କି ସାର୍। ବାରମ୍ବାର ଅଭିଯୋଗ ଜଲ୍ ପରେ ବୁଝି ପାରୁନାହିଁ , ଆଉ ଜେତେ ବର୍ଷ ଅଭିଯୋଗ ଜଲେ ବୁଝି ପାରିବେ ସାର୍ ଏ ବିଷୟରେ ଉତ୍ତର ଦେଇ ପାରିବେ କି।ମୁଁ ଜଣେ ଗରାଁବ ଯୁବକ ମୁଁ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ ପାଇଁ [redacted] ଡରଫଡୁ ଯୋଜନା ଆରମ୍ଭ ହୋଇ ଥିବା ଆବେଦନ କଲି ହେଲେ ବର୍ଷ ବର୍ଷ [redacted] କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଆରମ୍ଭ ହୋଇ ଅଛି କି ସାର୍ ଏ ବିଷୟରେ ଉତ୍ତର ଦେଇ ପାରିବେ କି ସାର୍।ଯଦି ମୋର୍ ଋଣ ମଞ୍ଜୁର ନେ ହୁଏ ତେହେଲେ ବ୍ୟାଙ୍କ୍ ଖାତା ସାର୍ନ୍ନ ରେ [redacted] ଦବିବାକୁ ବାଧ୍ୟ ହେବି।ମୋର୍ ହୋଇ ଥିବା କ୍ଷତି ପୂରଣ କିଏ ଭରଣ ପୋଷଣ କରିବ ମୁଁ ବ୍ୟବସାୟ ଆରମ୍ଭ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଆବେଦନ କଲି ହେଲେ ହୋଇସାର କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଆରମ୍ଭ ହୋଇ ଅଛି କି ସାର୍ ସରକାର ଶିକ୍ଷିତ ଯୁବକ ଓ ଯୁବତୀ ମାନଙ୍କୁ ଅଭିଆନ ତଥା ରୋଜଗାର [redacted] ପାଇଁ PMEGP ଯୋଜନା ଆରମ୍ଭ ହୋଇ ଥିଲେ ସୁଦ୍ଧା ଜାଗାର ଜଲମରେ ଅଟକି ଯାଇଥାଇଁ ।ମୁଁ ଦୁଇ ମାସ ହେବ ସମସ୍ତଙ୍କ ପ୍ରକାର ଜାଗାର ପତ୍ର ଦାଖଲ କଲା ପରେ ବି ସୁଦ୍ଧା ଋଣ ମଞ୍ଜୁର ହୋଇ ପାରୁନାହିଁ ଏହା ଦୁଃଖର ବିଷୟ ମୁଁ ଜଣେ ଗରାଁବ ଯୁବକ ବୋଲି ମୋର୍ ଋଣ ମଞ୍ଜୁର କରିବା ପାଇଁ ବର୍ଷ ବର୍ଷ ଲାଗୁଛି ହେଲେ ଜୋଟି ପୋଟି ଲୋକଙ୍କୁ ଖୁବ୍ ଶାଘ୍ର ଋଣ ମଞ୍ଜୁର କରାଯାଉଛି ।ଏଣୁ ଆପଣଙ୍କୁ ବିନୀତ ଅନୁରୋଧ ଜଲେ ଜି ଜାଗାର କଲମରେ ଅଟକି ରହିଥିବା ଋଣ କୁ ଯଥା ଶିଘ୍ର ବ୍ୟାଙ୍କ୍ କୁ ନିର୍ଦ୍ଦେଶ ଦେଇ ଖୁବ୍ ଶାଘ୍ର ଋଣ ମଞ୍ଜୁର କରା ଗଲେ ମୁଁ ଆପଣଙ୍କ ନିଜଟରେ ଚିର ଉପକୃତ ହେବି।
DEABD/E/2023/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted]Branch / Name of Bank and Branch : [redacted]Date of Application [redacted]----------To, The [redacted] [redacted]ବିଷୟ: [redacted] ଫିଲ୍ଡ ଅଫିସର ନେ ମୁଝେ PMEGP ଋଣ ନେହି ଦେନେ କେଲିଏ ବରସେ ହୁए मेरो सब डॉक्यूमेंट पेंडिंग रखा ओर ट्रैक टरी किया ओर मेरी पचास हजार का मानहानि ख्याति किया मुझे मेरे PMEGP ऋण केलिए परिसानी हु, इसलिए मैं आपको माफी मागता हूं की मुझे बार बार परिसानी किया इसलिए मैं बार बार अभियोग दर्ज करने केलिए मन किया, मे आपको अनुरोध किया हे [redacted] स्टेट ब्यांक का फील्ड ऑफिसर को मेरो बोध जिला से दूसरे जिलाको बदली करने केलिए अभियोग दर्ज किया मुन्ना भोई हे।Dear Sir/Madam, [redacted] का फील्ड ऑफिसर ने मुझे PMEGP ऋण प्रदान करने केलिए मेरेनू [redacted] का लांच लिया PMEGP ऋण प्रदान करने केलिए ओर आज होगी काल होगी बलते ओर मेरे पैसा लूटे हुए फील्ड ऑफिसर को मेरी [redacted] से डुसेरा जिला को बदली करने केलिए आपको बिनित अनुरोध किया हे जल्द से जल्द फील्ड ऑफिसर को बदली कीजे मेरी जिला से [redacted] का फील्ड ऑफिसर मेरी मानहानि ख्याति किया ओर मेरे PMEGP ऋण में ट्रैक टरी किया ओर सब डॉक्यूमेंट ब्रांच मे पेंडिंग रखा फील्ड ऑफिसर को मेरी जिला में रहेने मे जरूरत नेही हे , सरकार का कोटि कोटि टंका लूटा फील्ड ऑफिसर को तुरंत बदली कीजे।
DEABD/E/2023/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted]Branch / Name of Bank and Branch : [redacted]Date of Application : [redacted]----------To, The [redacted]Sub: [redacted] branch field officer and branch manager posting transfer others district.Submitted by to the branch [redacted] all documents and one year ago my loan pending our [redacted] branch field officer my loan pending [redacted] compliance others district field officer and branch manager my [redacted] district not allowed immediately posting transfer others district.Any time margin money was deposit the [redacted] branch PMEGP scheme under and any time my loan rejected [redacted] branch field officer [redacted] treated my [redacted] district not possible others district officer immediately whatever cancle field officer and branch manager posting.my PMEGP loan application Racc [redacted] not forward the [redacted] field my documents pending one year ago our [redacted] branch and [redacted] field field officer rejected any time may bi treated govt PMEGP loan scheme under any time posting transfer compliance. this PMEGP loan proposed I am invest to [redacted] above after [redacted] running to the [redacted] branch to PMEGP scheme under and running to the DIC office boudh any time PMEGP loan scheme under and [redacted] investment PMEGP scheme loan under immediately my payment refund [redacted] branch field officer my payment immediately refund.Dear sir my pending loan immediately speedly processing Nehi ta any time compliance process [redacted] action the my compliance thanks u so much.
MMSME/E/2022/[redacted]	Micro Small and Medium Enterprises >> Schemes >> PMEGPApplicant Details : Pmegp loan scheme compliance----------To, The Msme pmegp government of India departments [redacted] Dear sir/MadamDear sir one kind any time all documents submitted our [redacted] branch and [redacted] branch field officer one kind loan application any time rejected i am be treated [redacted] branch field officer se government PMEGP loan scheme me i am be any time compliance government PMEGP scheme me.Dear sir request your kindly and contact the connar agency [redacted] and re forward the one kind pmegp loan application [redacted] [redacted] branch please sir one'kind loan problem solved immediately action.One kind any time running the [redacted] branch and [redacted] running any time and contact the agency KVIB officer boudh re forward the loan application [redacted] branch one kind all documents submitted our [redacted] branch and night months ago one kind all documents pending and field officer one kind loan not sanction and CIBIL score low field officer one kind pmegp loan and other application pmegp loan sanction CIBIL score high score field officer and immediately others pmegp loan application process for sanction one kind loan pending one kind EDP traning under and skills india training under and udyog aadhar certificate under and income tax under and GST certificate under and one kind Pmegp loan not sanction [redacted] [redacted] branch field officer.any time [redacted] branch field officer one kind Pmegp loan application rejected immediately [redacted] branch field officer transfer others district immediately action my compliance please sir. [redacted] branch field officer contact the connar agency [redacted] and re forward the one kind Pmegp loan application immediately action please sir and all day compliance closed.
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank : [redacted]Branch / Name of Bank and Branch : [redacted]Date of Application : [redacted]----------To, The [redacted] [redacted]ବିଷୟ: [redacted] मे फील्ड ऑफिसर ने मुझे PMEGP ऋण नेही देने केलिए बरसे हुए मेरो सब डॉक्यूमेंट पेंडिंग रखा ओर ट्रैक टरी किया ओर मेरी पचास हजार का मानहानि ख्याति किया मुझे मेरे PMEGP ऋण केलिए परिसानी हु, इसलिए मैं आपको माफी मागता हूं की मुझे बार बार परिसानी किया इसलिए मैं बार बार अभियोग दर्ज करने केलिए मन किया, मे आपको अनुरोध किया हे [redacted] [redacted] का फील्ड ऑफिसर को मेरो बोध जिला से दूसरे जिलाको बदली करने केलिए अभियोग दर्ज किया मुन्ना भोई हे।Dear Sir/Madam, [redacted] का फील्ड ऑफिसर ने मुझे PMEGP ऋण प्रदान करने केलिए मेरेनू पचास हजार का लांच लिया PMEGP ऋण प्रदान करने केलिए ओर आज होगी काल होगी बलते ओर मेरे पैसा लूटे हुए फील्ड ऑफिसर को मेरी [redacted] से [redacted] को बदली करने केलिए आपको बिनित अनुरोध किया हे जल्द से जल्द फील्ड ऑफिसर को बदली कीजे मेरी जिला से [redacted] [redacted] का फील्ड ऑफिसर मेरी मानहानि ख्याति किया ओर मेरे PMEGP ऋण में ट्रैक टरी किया ओर सब डॉक्यूमेंट ब्रांच मे पेंडिंग रखा फील्ड ऑफिसर को मेरी जिला में रहेने मे जरूरत नेही हे , सरकार का कोटि कोटि टंका लूटा फील्ड ऑफिसर को तुरंत बदली कीजे।

Figure 3.2: Sampled List of Grievances - 2. Tokens that could reveal personal identifiable information (PII) have been redacted.

3.5.2.2 Captured Bulk Grievances

reg_no	subject_content
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank :[redacted] / Name of Bank and Branch :[redacted] of Application :[redacted]----------To, The[redacted]ବିଷୟ: ଜଣ୍ଟୋମାଳ ବ୍ଲଜର ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ [redacted] ଖଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସର ଲୌଣସି ଜାମ କରୁ ନୋ ଥିବାରୁ ଫିଲ୍ଡ ଅଫିସରଙ୍କୁ ଖୁବ୍ ଶୀଘ୍ର ବଦଲି କରାଯାଇ ଜଣ୍ଟୋମାଳ ବ୍ଲଜର ଗରୀବ ଯୁବକ ମାନଙ୍କ ଜରିଆରେ ଫିଲ୍ଡ ଅଫିସର ବଦଲୀ ସକାଶେ ମାନ୍ୟବର, ଲେଖି ଜଣାଇବାର କାରଣ ଏହି କି ଯେ ବିଗତ ଶହ ଜଣ୍ଟୋମାଳ ବ୍ଲଜର ଗରୀବ ଯୁବକ ମାନେ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବାରମ୍ବାର ଫିଲ୍ଡ ଅଫିସର ରିଯେଷ୍ଟ କରିଥିବାରୁ ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ଜାହିଁକି ଫିଲ୍ଡ ଅଫିସର ବଦଲି କରା ଯାଉ ନାହିଁ। ଫିଲ୍ଡ ଅଫିସର [redacted]ପାଉ ଧାନ ବିକ୍ରୀ କରି ସରକାରୀ PMEGP ଋଣ ପ୍ରଦାନ କରୁଛନ୍ତି [redacted] ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବିଭିନ୍ନ ପ୍ରକାର ନାଟ ଦେଖାଇ ଗରୀବ ଯୁବକ ମାନଙ୍କ CIBIL score low କରି ବାରମ୍ବାର ରିଯେଷ୍ଟ କରିଥିବାରୁ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଲାଗର ପତ୍ରକୁ ଦାଖଲରେ ଖର୍ଚ୍ଚ ହୋଇ ଥିବା ଟଙ୍କା ବ୍ୟାଙ୍କ ଭାବେ ମାନହାନି କ୍ଷତି କରିଥିବା [redacted]ଅଫିସର [redacted] ଖଟେଟ [redacted] ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ ତୁରନ୍ତ ବଦଲି ଦାବିରେ ।ଅନ୍ୟ ଲୋକଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ CIBIL score ବଢ଼ାଇ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି PMEGP ଋଣ ପ୍ରଦାନ କରୁଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କ CIBIL ସ୍କୋରରେ ନାଟ କରି ସମସ୍ତ ପ୍ରକାର ଜାଗାଜ ପତ୍ରର ବର୍ଷ ବର୍ଷ ଧରି pending ରଖି ଆଜି ଦେଖିବା ଜାଲି ଦେଖିବା କହି ବାରମ୍ବାର ରିଜେଷ୍ଟ କରୁଥିବା ଫିଲ୍ଡ ଅଫିସର ବୈଦ୍ୟ ଜିଲ୍ଲାର [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ରହିବାର ଆବଶ୍ୟକତା ତୁରନ୍ତ ବଦଲି କରାଯାଉ ବୋଲି ଜଣ୍ଟୋମାଳ ବ୍ଲଜର ଗରୀବ ଯୁବକ ମାନେ ଦାବି ପତ୍ର ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ସରକାରୀ ଅଫିସର୍ ଗଣ ଜାହିଁକି ବଦଲି କରୁନାହିଁନ୍ତି ଜବାବ୍ [redacted] ।PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରୁ କିଛି ଅଫିସର୍ ମାନେ ବ୍ୟାପାର କରି ମଦ ମାଂସ ଖାଇ ଜମିଶନ ନେଉ ଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନେ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ସମସ୍ତ ତଥ୍ୟ ଗୁଡ଼ିକ ପ୍ରଦାନ କରିବା ପାଇଁ ସବୁ ତଥ୍ୟ ଗପି ରଖିଛନ୍ତି । PMEGP ଋଣ ଯୋଜନା ଉପରେ [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ଚୁଟ୍ ଚଣ୍ଡାଳ ମାନଙ୍କୁ PMEGP ଋଣ ପ୍ରଦାନ କରୁଥିବା ବେଳେ ସମସ୍ତ ପ୍ରକାର ଲିଖିତର ତଥ୍ୟ PMEGP ପୋର୍ଟାଲ ମଧ୍ୟ୍ୟମାରେ ପରିଶୋଟ କରି ପ୍ରଦାନ କରିପାରୁ ନେଥିବା ବେଳେ ବ୍ୟାପକ ଭାବେ ଲୋଟି ଲୋଟି ଟଙ୍କା ଘୁମାଁଟି କରି ସମସ୍ତ ତଥ୍ୟ ଗପା ରଖୁଛନ୍ତି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସର ।ଏଣୁ ଆପଣଙ୍କୁ ବିନୀତ ଅନୁରୋଧ କରୁଛି ଯେ ଉପରୋକ୍ତ ବିଷୟ ପ୍ରତି ଧ୍ୟାନ ଦେଇ ଖୁବ୍ ଶୀଘ୍ର ପଦକ୍ଷେପ ନିଆଯାଉ ।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank [redacted]Branch / Name of Bank and Branch :[redacted]of Application :[redacted]----------To, The[redacted]ସରକାର ଜାହିଁକି ଗରିବ ଯୁବକ ଓ ଯୁବତୀ, ଜଣା ସାଧାରଣ ଲୋକ ମାନଙ୍କ ପାଇଁ PMEGP ଋଣ ଯୋଜନା ପ୍ରାସ୍ତୁତ କରି ଛନ୍ତି, ସରକାରୀ ଯୋଜନାରେ ସାମିଲ ହୋଇ ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ନା ବର୍ଷ ବର୍ଷ ଧରି ଆଶା ଦେଇ ସମସ୍ତ ପ୍ରକାର ଜାଗାଜ ପତ୍ର ପଜାଇ ରଖି ହୋଇ ଋଣ କରିବା ପାଇଁ [redacted] PMEGP ଯୋଜନା ପ୍ରାସ୍ତୁତ କରିଛନ୍ତି ।ଏ ବିଷୟରେ ବାବୁ ଅଫିସର ମାନେ ଟିକେ ଉତ୍ତର ଦେବେକି ।ସରକାର ଫିଲ୍ଡ ଅଫିସର ଏବଂ ବ୍ୟାଙ୍କ [redacted] ଜାହିଁକି ନିଯୁକ୍ତି ଦେଇ ଛନ୍ତି, [redacted] ଯୁବକ ଓ ଯୁବତୀ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ଋଣ ଆବେଦନ କଲେ ତମକୁ ଅଫିସର ମାନଙ୍କୁ ଜାହିଁ ଭିତ୍ ପତ୍ର ଛେ ବ ତୁମର ଦରମା ଟଙ୍କାର PMEGP ଋଣ [redacted]କରୁଛି କି ଯେ ଏତେ ଭିତ୍ ପତୁଛେ ।ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାର ସାହାୟ୍ୟ କରୁଥିଲେ ମଧ୍ୟ ଜଟକିଆ ଅଫିସର ମାନେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ଲୁଟିଗଲି ଛନ୍ତି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ସରକାରୀ PMEGP ଋଣ ପ୍ରଦାନ ହେବା ପାଇଁ, ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ବୈଦ୍ୟ ଜିଲ୍ଲାରେ ସରକାରି ଯୋଜନାରେ ଲୋଟି ଲୋଟି ଟଙ୍କା ଲୁଟି ଅନ୍ୟ ଲୋଟି ପୋଟି ଲୋକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି ମଦ ମାଂସ ଭୋଜି ଖାଇ ଜାମା କରି ଗାଲୁ ଥିବା ବେଳେ ଗରିବ ଯୁବକ ମାନଙ୍କୁ ବହୁତ୍ ଅନ୍ୟାୟ କରି ଗଲିଥିବା ଅଫିସର ମାନଙ୍କୁ ଓସେମ୍ବର ମାସ ସୁଦ୍ଧା ବଦଲି ଦରକାରୁ ।ଲୋଟି ଲୋଟି ଟଙ୍କା ଲୁଟୁ ଥିବାରୁ PMEGP ପୋର୍ଟାଲ ସାଇଟ୍ ରୁ PMEGP ଋଣ ଆବେଦନ କରିଥିବା ସମସ୍ତ ତଥ୍ୟ ଦେଇ ପାରୁ ନାହିଁ, ବାରମ୍ବାର ମିଥ୍ୟା ତଥ୍ୟ ଦେଇ ଗଲି ଛନ୍ତି ତୁରନ୍ତ ସମସ୍ତ ଲିଖିତ ଆବଶ୍ୟକ [redacted] ମସିହା ଠାରୁ [redacted] ମସିହା ପର୍ଯ୍ୟନ୍ତ ଋଣ ନେଇ ଥିବା ଆବେଦନ ଜାଗାଜ୍ଙ୍କ ତଥ୍ୟ ଦରକାର୍।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank :[redacted] Name of Bank and Branch :[redacted]Date of Application [redacted]----------To, The[redacted]finance minister department government of India [redacted]ବିଷୟ: [redacted] ଖଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସରଙ୍କୁ ନିର୍ଦ୍ଦେଶ ଦିଅନ୍ତୁ ଯେ ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ରିଯେଷ୍ଟ କରିଥିବା ପୁନର୍ବାର [redacted]ଖଟେଟ ବ୍ୟାଙ୍କକୁ re forward କରି ଗରୀବ [redacted] ମାନଙ୍କ ଋଣ ପ୍ରଦାନ କରା ଯାଉ ନୋ ହେଲେ ଆଗକୁ ଆହୁରି ଅନେକ ଅଭିଯୋଗ ଦାଖଲ ଜାଗା ରହିବ ଫିଲ୍ଡ ଅଫିସରଙ୍କ ବିରୁଦ୍ଧରେ ଦୃଢ଼ କାର୍ଯ୍ୟାନୃଷ୍ଠାନ ଗ୍ରହଣ କରାଯାଉ । ଫିଲ୍ଡ [redacted] ଯଦି [redacted]ଖଟେଟ ବ୍ୟାଙ୍କରେ ସବୁ ଦିନ ଜାମ କରିବା ପାଇଁ ଗଡୁଛି ଉହାଲେ ଗରୀବ ଯୁବକ [redacted] ଋଣ ପ୍ରଦାନ କରା ଯାଉ ନୋ ହେଲେ ଅଭିଯୋଗ ସବୁ ଦିନ ପାଇଁ ଜାରି ରହିବ ।ମାନ୍ୟବର, ଲେଖି ଜଣାଇବାର କାରଣ ଏହି କି ଯେ ବୈଦ୍ୟ ଜିଲ୍ଲାର ସ୍ୱାୟୀ ବାସିନ୍ଦା ହୋଇ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ପାଇଁ ଋଣ ପ୍ରଦାନ ହୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ନାଟ କରି ସରକାରୀ ଯୋଜନା ଅନୁସାରେ [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଋଣ ପ୍ରଦାନ ହୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ନାଟ କରି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ଥିବା ଫିଲ୍ଡ ଅଫିସର ସରକାରୀ PMEGP ଋଣ ପ୍ରଦାନ [redacted] ବଡ଼ ବଡ଼ଆ ଲୋକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରି ଋଣ ପ୍ରଦାନ କରୁଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ବର୍ଷ ବର୍ଷ ଧରି ଆଶ୍ୱସନା ଦେଇ ଆଜି ଦେଖିବା ଜାଲି ଦେଖିବା କହି ସମସ୍ତ ପ୍ରକାରର ଜାଗାଜ ପତ୍ରର ବର୍ଷ ବର୍ଷ ଧରି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ପଜାଇ ରଖୁଥିବା ଫିଲ୍ଡ ଅଫିସର [redacted]ଖଟେଟ ବ୍ୟାଙ୍କରେ ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ, ଅନ୍ୟ ଲୋକଙ୍କ [redacted]ପରବାର କରିବା ପାଇଁ ଖୁବ୍ ଶୀଘ୍ର ଜାମ ଆରମ୍ଭ କରୁ ଥିବା ବେଳେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଜାମ ନେ କରୁଥିବା ଫିଲ୍ଡ ଅଫିସର ବୈଦ୍ୟ ଜିଲ୍ଲାରୁ ତୁରନ୍ତ ବଦଲି କରାଯାଉ ।CIBIL ସ୍କୋରରେ ବିଭିନ୍ନ [redacted] ନାଟ ଦେଖାଇ ଗରୀବ ଯୁବକ ମାନଙ୍କ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବାରମ୍ବାର ରିଯେଷ୍ଟ କରୁଥିବା ବେଳେ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ [redacted] ପ୍ରତିଷ୍ଠାନ କରିବା [redacted]ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ଋଣ ପ୍ରଦାନ ହୋଇ ପାରୁନାହିଁ, ସରକାର PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ବ୍ୟାଙ୍କ ଗୁଡ଼ିକୁ ନିର୍ଦ୍ଦେଶ ଦେଇଥିଲେ ମଧ୍ୟ ସରକାରଙ୍କ କଥାକୁ ଫୁଂ୍ ଉଡୁଥିବା ଅଫିସର ବୈଦ୍ୟ [redacted] ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ ।ମାନ୍ୟବର, ଲେଖି ଜଣାଇବାର କାରଣ ଏହି [redacted] ବୈଦ୍ୟ ଜିଲ୍ଲାର ସ୍ୱାୟୀ ବାସିନ୍ଦା ହୋଇ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ସ୍ୱବାବଲମ୍ବୀ ହେବା ପାଇଁ ଜାହିଁକି PMEGP ଋଣ ପ୍ରଦାନ ହୋଇ ପାରିବ ନାହିଁ, ସରକାର PMEGP ଋଣ ଯୋଜନା ଜାହିଁକି ପ୍ରସ୍ତୁତ କରିଛନ୍ତି, ଋଣ ନେଇ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ନା ବର୍ଷ ବର୍ଷ ଧରି ଆଶ୍ୱସନା ଦେଇ ସମସ୍ତ ପ୍ରକାର ଜାଗାଜ ପତ୍ରର ପେଣ୍ଡିଡ୍ଙ୍ଗ ରଖି ହୋଇସାର କରିବା ପାଇଁ ସରକାର ଯୋଜନା ପ୍ରସ୍ତୁତ କରିଛନ୍ତି କି ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ବୈଦ୍ୟ ଜିଲ୍ଲାରେ ଅଫିସର୍ ମାନେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଉପରେ ଅନ୍ୟାୟ କରୁଥିବା ବେଳେ ବାରମ୍ବାର [redacted]ଜଲେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ମିଳି ପାରୁନାହିଁ, ଆଉ କେତେ ବର୍ଷ ଅଭିଯୋଗ [redacted] ନ୍ୟାୟ ମିଳି ପାରିବ ଏ ବିଷୟରେ ଉତ୍ତର ପ୍ରଦାନ କରାଯିବ କି, ଗରୀବ ଯୁବକ ମାନେ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ ବାରମ୍ବାର ଅଭିଯୋଗ ଜଲେ ବୁଡି ପାରୁ ନାହିଁ ଆଉ କେତେ ବର୍ଷ ଅଭିଯୋଗ ପରେ ବୁଝି ପାରିବେ ସାର୍।
DEABD/E/2022/[redacted]	Financial Services (Banking Division) >> Government sponsored Schemes Related >> Prime Minister Rojgar YojanaBank :[redacted] / Name of Bank and Branch :[redacted]Date of Application :[redacted]----------To, The[redacted]ବିଷୟ: [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ସରକାରୀ PMEGP ଯୋଜନା ଉପରେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଉପରେ ବ୍ୟାପକ ଭାବେ ଅନ୍ୟାୟ କରୁଥିବା ବେଳେ ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ସରକାରୀ ଯୋଜନା ଅନୁସାରେ ଗରୀବ [redacted] ମାନଙ୍କୁ ନ୍ୟାୟ ପ୍ରଦାନ ହୋଇ ପାରୁନାହିଁ, ଆଉ [redacted] ବର୍ଷ ଅଭିଯୋଗ କଲେ ଗରୀବ ଯୁବକ ମାନେ ନ୍ୟାୟରୁ ବଞ୍ଚି ପାରିବେ , [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ ଋଣ [redacted] ପାଇଁ ବାରମ୍ବାର [redacted]କରିଥିବା ଫିଲ୍ଡ ଅଫିସର ବୈଦ୍ୟ ଜିଲ୍ଲାରେ ରହିବାର ଆବଶ୍ୟକତା ନାହିଁ ତୁରନ୍ତ ବଦଲି ଦାବି ।ମାନ୍ୟବର, ଲେଖି ଜଣାଇବାର କାରଣ ଏହି କି ଯେ ବୈଦ୍ୟ ଜିଲ୍ଲାର ସ୍ୱାୟୀ ବାସିନ୍ଦା ହୋଇ ଗରୀବ ଯୁବକ ମାନେ ବ୍ୟବସାୟ ପ୍ରତିଷ୍ଠାନ କରିବା ପାଇଁ [redacted]ଯୋଜନା ଅନୁସାରେ ଋଣ ପ୍ରଦାନ ହୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ନାଟ କରି ଜାଗାଜ ଜଲମରେ ଅଟକାଇ ରଖିଥିବା ଋଣ ଖୁବ୍ ଶୀଘ୍ର ପ୍ରଦାନ କରାଯାଉ ନୋ ହେଲେ [redacted]ସବୁ ଦିନ ପାଇଁ ଜାଗା ରହିବ ।ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ଅଫିସର ମାନେ [redacted] ଜିଲ୍ଲାର ଗରୀବ ଯୁବକ ମାନଙ୍କ ଉପରେ ବହୁତ୍ ଅନ୍ୟାୟ କରୁଥିବାରୁ ବାରମ୍ବାର ଅଭିଯୋଗ କଲେ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ପ୍ରଦାନ ହୋଇ ପାରୁନାହିଁ ବର୍ଷ ବର୍ଷ ଧରି ବିଭିନ୍ନ ପ୍ରକାର ନାଟ କରି ଗଲିଛନ୍ତି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କର ଫିଲ୍ଡ ଅଫିସର ।ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି ବୈଦ୍ୟ ଜିଲ୍ଲାର ଜଣାଟୋମାଳ ବ୍ଲଜର [redacted] ଖଟେଟ ବ୍ୟାଙ୍କରେ [redacted] ମାନଙ୍କ PMEGP ଋଣ ଯୋଜନା ଉପରେ ବହୁତ୍ ଅନ୍ୟାୟ [redacted] ବାରମ୍ବାର ଅଭିଯୋଗ ଦାଖଲ କରୁଥିଲେ ମଧ୍ୟ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ପ୍ରଦାନ ମିଳି [redacted] ତୁରନ୍ତ ସମାଧାନ ଦାବି ନୋ ହେଲେ ଅଭିଯୋଗ ସବୁ ଦିନ ପାଇଁ ଜାଗା ରହିବ ।ଲୋର ଅଫିସରଙ୍କୁ ଓରେଲ PMEGP ଋଣ ପ୍ରଦାନ କରିବା ପାଇଁ ଜହୁମାନ୍ତ୍ର ବୈଦ୍ୟ ଜିଲ୍ଲାର ଜଣାଟୋମାଳ ବ୍ଲଜର ଗରୀବ ଯୁବକ ମାନେ, ଫିଲ୍ଡ ଅଫିସର ଗରୀବ ଯୁବକ ମାନଙ୍କ ବ୍ୟାପକ ଭାବେ ମାନହାନି କ୍ଷତି କରି ବାରମ୍ବାର PMEGP ରିଜେଷ୍ଟ କରିଥିବାରୁ [redacted] ଯୁବକ ମାନଙ୍କ ଅଧିକାର ଛଡେଇ ନେଉ ଥିବା ଫିଲ୍ଡ ଅଫିସର ବହୁତ୍ ଅନ୍ୟାୟ କରୁଛନ୍ତି ।ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ଜେଲ୍ ଫାଶୀ ହେଇ ଯାଉ ପଛେ ଗରୀବ ଯୁବକ ମାନଙ୍କ ଜୀବନ ପଛେ ଗଲିଯିବ ସରକାରୀ PMEGP ଯୋଜନା ଉପରେ ଲଢ଼େଇ ଜାରି ରହିବ ଯେ ପର୍ଯ୍ୟନ୍ତ ଗରୀବ ଯୁବକ ମାନଙ୍କୁ ନ୍ୟାୟ ନୋ ମିଳିଛି ସେ ପର୍ଯ୍ୟନ୍ତ ଅଭିଯୋଗ ଜାରି ରହିବ ।ଦେଖା ଯାଉ କେତେ କେତେ ବଡ ବାଡୁ ଆ ଫିଲ୍ଡ ଅଫିସର ମାନେ ଅନ୍ୟ ଜିଲ୍ଲାରୁ ଆସି [redacted] ଖଟେଟ ବ୍ୟାଙ୍କକୁ ଗଜିରି କରିବାକୁ ଆସୁଛନ୍ତି ଦେଖା ଯାଉ ଅଭିଯୋଗ ଜାରି ରହିବ ।ଜଣ୍ଟୋମାଳ ବ୍ଲଜର ଗରୀବ ଯୁବକ ମାନଙ୍କ ଜରିଆରେ ଚେତାବନୀ।

Figure 3.3: Captured Bulk Grievances. Tokens that could reveal personal identifiable information (PII) have been redacted.

3.6 Conclusion and Future Work

In this chapter, we proposed an approach to efficiently cluster multilingual text data using pre-trained word embeddings and hierarchical navigable small world (HNSW) indexing. We first applied a text cleaning function to remove any unwanted characters or noise from the input text data. We then generated 300-dimensional vector representations of each text using pre-trained embeddings dictionaries for nine different languages. These embeddings were used to fit an HNSW index with an exploration factor of K*10, where K is set to the minimum of the length of the DataFrame and 1000.

Next, we found the K nearest neighbors and their distances for each embedding and calculated the L2 distance threshold to identify clusters of similar documents. We then created a bulk_all list that contained all the similar documents within the L2 distance threshold for each primary document. The primary documents were then assigned a representative document ID and the length of their respective bulk using the output dictionary.

As future work, we could explore the use of different clustering algorithms and evaluate their performance on multilingual text data. Additionally, we could investigate the use of different pre-processing techniques to further improve the quality of the embeddings. Finally, we could explore the application of our approach to real-world scenarios such as topic modeling, sentiment analysis, and recommender systems.

Chapter 4

Multilingual Semantic Search

Contents

Abstract *This chapter presents a system for language detection, text cleaning, translation, and similarity search. It involves reading a dictionary for word standardization, defining Unicode ranges for languages, and implementing functions for language detection and text cleaning. Pre-trained models handle translation and sentence embedding. An index is created using the "hnswlib" package, utilizing translated English texts and sentence embeddings. This work enables efficient analysis of multilingual data, contributing to language processing and information retrieval.*

4.1 Introduction

In today's data-driven world, the ability to extract valuable insights from large-scale datasets is essential for informed decision-making. When it comes to public grievances, the stakes are high, as the efficient resolution of these issues is crucial for maintaining public trust and confidence in government institutions. In this context, I am pleased to present my thesis on a cutting-edge semantic engine designed to process vast amounts of public grievances data, providing government officials with actionable insights to improve service delivery and increase citizen satisfaction. By leveraging the latest advancements in natural language processing and machine learning, this system has the potential to revolutionize the way governments respond to public grievances, leading to a more responsive, efficient, and effective public service.

4.2 Problem Statement

Create a search system that is semantic and can be used across large datasets that supports multiple languages.

4.3 Related Work

Several studies have been conducted in the area of public grievance redressal systems, highlighting the need for effective and transparent mechanisms to address citizen complaints. Choudhary et al. (8) propose a machine learning-based approach to predict the priority level of grievances in India, emphasizing the importance of using technology to manage public grievances. Joshi et al. (10) present the design and implementation of an online public grievance redressal system for Indian government agencies, stressing the need for efficient and transparent grievance redressal mechanisms. Rahman et al. (17) propose a sentiment analysis approach to identify areas for government services improvement from citizen complaints. Ahmed et al. (5) propose a citizen-centered framework for public service delivery that integrates citizen feedback into the service delivery process. Finally, Saha et al. (19) propose a social media analytics-based approach to improve the responsiveness and effectiveness of the public grievance redressal system in India. These studies provide insights into the various approaches and techniques that can be used to improve public grievance redressal systems.

4.4 Methodology

1. The first step is to read a file "final_wording.pkl" that contains a dictionary with certain key-value pairs. The dictionary has mapping between certain words and their standardized version. This is done using the Python Pickle module which can serialize and deserialize complex data structures.

2. Next, multiple Unicode ranges for different languages such as Hindi, English, Punjabi, Gujarati, Telugu, Tamil, Kannada, Odia, and Bengali are defined. These ranges contain the Unicode code points for the characters present in each language.

3. A function named "detect_language" is defined that takes a sentence as input and counts the number of characters present in each language range. The language with the highest count is considered to be the language of the sentence and returned.

4. Another function named "clean_text" is defined that takes a string as input and performs certain text cleaning operations such as removing URLs, punctuation marks, digits, special characters, extra whitespaces, etc. The function also replaces certain words in the input string with their standardized versions from the dictionary loaded earlier.

5. The next step involves reading a CSV file "chunk3.csv" that contains data related to a subject and its corresponding registration number. The data is read into a

Pandas dataframe and the required columns are selected. The ”clean_text” function is applied to the ”subject_content” column to clean the text. The resultant cleaned text is stored in a new column ”subject_content_cleaned”. The rows containing missing or empty ”subject_content_cleaned” values are removed. The ”subject_content_cleaned” column is then split into multiple rows using ”explode” function.

6. A pre-trained multilingual model ”MBartForConditionalGeneration” is loaded from the ”facebook/mbart-large-50-many-to-one-mmt” checkpoint and moved to the specified device (e.g., CPU or GPU). A tokenizer for the same model is also loaded from the ”facebook/mbart-large-50-many-to-one-mmt” checkpoint.

7. A pre-trained sentence embedding model ”SentenceTransformer” is loaded from the ”salesken/similarity-eng-hin_latin” checkpoint and moved to the specified device.

8. A function named ”translate_batch” is defined that takes a batch of input texts, the source language of the texts, and the batch size as inputs. If the source language is ”English”, the function returns the input texts as it is. Otherwise, the function uses the MBART model and tokenizer to translate the input texts from the source language to English. The function returns the translated texts.

9. Finally, an index for a similarity search is created using the ”hnswlib” package. If an index file already exists, it is loaded, else a new index is created. The data stored in the Pandas dataframe is used to build the index. Each subject is translated to English (if not already in English) and its sentence embeddings are computed using the ”SentenceTransformer” model. The embeddings are then added to the index along with the corresponding registration number.

4.5 Algorithms

4.5.1 HNSW

Algorithm 2 HNSW algorithm

0: Initialize the graph G with the first data point x_1.
0: **for** $i = 2$ to n **do**
0: Find the nearest neighbor x_{nn} of x_i in the graph G.
0: Add x_i as a new node to the graph G.
0: Connect x_i to x_{nn} in the graph G.
0: **for** each level l in the graph G **do**
0: Find the approximate nearest neighbor x_{ann} of x_i at level l.
0: Connect x_i to x_{ann} at level l in the graph G.
0: **end for**
0: **end for**=0

This pseudocode assumes that the input data is a set of n points, and that the graph G is an HNSW graph, which is a hierarchical graph where each level corresponds to a different resolution of the data space. At each level, the graph contains a set of nodes and edges connecting them. Each node represents a data point, and each edge represents a connection between two nodes.

The algorithm starts by initializing the graph G with the first data point x_1. Then, for each subsequent data point x_i, it finds the nearest neighbor x_{nn} of x_i in the graph G, and adds x_i as a new node to the graph G, connecting it to x_{nn}. It also connects x_i to its approximate nearest neighbors at each level in the graph G. This process is repeated for all n data points, resulting in a fully connected HNSW graph.

4.6 Experiments

4.6.1 Comparing Language Models for English-Hindi Sentence Similarity and Evaluation on the Hinglish-TOP Dataset

In this experiment, a comparison was made between two language models, namely the SentenceTransformer model "salesken/similarity-eng-hin_latin" and the Hinglish language model "l3cube-pune/hing-bert". The purpose of the comparison was to determine which model is better suited for measuring the similarity between English and Hindi sentences.

This experiment also utilized a large annotated dataset called Hinglish-TOP. This dataset consists of two parts: a human annotated code-switched semantic parsing dataset containing 10,000 examples. (20)

The human annotated code-switched semantic parsing dataset was created by annotating queries derived from the TOPv2 multi-domain task oriented semantic parsing dataset. This dataset is unique in that it contains code-switching between Hindi and English, which is a common phenomenon in India. The annotations provide a mapping between the natural language input and the structured output representation.

The first step involved cleaning the data by removing any unwanted characters or symbols from the English and Hindi sentences. Then, a dataset of 1000 pairs of English and Hindi sentences was selected for analysis.

Two functions were created to calculate the similarity scores between the English and Hindi sentences using the SentenceTransformer and Hinglish language models, respectively. The "compare_using_salesken" function used the SentenceTransformer model "salesken/similarity-eng-hin_latin", while the "compare_using_hingbert" function used the Hinglish language model "l3cube-pune/hing-bert". Both functions calculated the cosine similarity between the embeddings of the English and Hindi sentences.

The results were analyzed by calculating the mean similarity scores for each model. The SentenceTransformer model "salesken/similarity-eng-hin_latin" had a mean similarity score of **0.982997031**, while the Hinglish language model "l3cube-pune/hing-bert" had a mean similarity score of **0.860918048**.

In conclusion, the comparison between the SentenceTransformer and Hinglish language models revealed that both models are effective in measuring the similarity between

English and Hindi sentences. However, the results showed that the the SentenceTransformer model ”salesken/similarity-eng-hin_latin” performed slightly better than Hinglish language model ”l3cube-pune/hing-bert” in terms of mean similarity score.

4.7 Results

1. Query: Air and water pollution harm health

reg_no	subject_content	Language
DARPG/P/2019/[redacted]	Reg. Poor sanitation and dengue death.	English
DARPG/E/2019/[redacted]	[redacted] Municipal Corporation in [redacted] is developing a scientific landfill site at [redacted] which is now a prime\nlocation of the city. The site is just [redacted] meter from a residential society having [redacted] flats. Also there are\nmore than [redacted] flats in the vicinity of the site and a government school within [redacted]. It is also on\nthe bank of River [redacted] which flooded the entire area including the site during recent deluge. This site is in\ntotal violation of [redacted] / [redacted] norms. These facts have been ignored or falsely put in [redacted] / site selection report. The citizens have opposed this at the time of public hearing but the objections have been ignored or\nfalsely put in the report. Request your intervention for reassessment of the site in the residential area and avoiding health hazard to the residents.	English
DHLTH/E/2021/[redacted]	Health & Family Welfare >> Public Health >> Disease Outbreak >> Other ---------- [redacted] ડાઇંગ ઉદ્યોગનું કેમિકલ વાળું પાણી [redacted] જુલ્લાના દરિયામાં તેમજ અન્ય કોઇપણ જગ્યાએ ના ઠાલવવા બાબત. જય [redacted] સાથ જણાવવાનું કે [redacted] જુલ્લામાં [redacted] નું ડાઇંગ ઉદ્યોગનું પાણી\n[redacted] ના દરિયામાં [redacted] થી [redacted] સુધી પાઇપ લાઈન મારફત સુદ્ધ કરીને દરિયામાં નાખવાનો જે [redacted] [redacted] નો પ્રોજેક્ટ આપશ્રી દ્વારા\n[redacted] સ્થિત [redacted]. લીમીટેડને આપવામાં આવ્યો છે તે પ્રક્રિયા તાત્કાલિક અસર થી બંધ કરી ને પ્રોજેક્ટ કેન્સલ\nકરવા અરજ છે. [redacted] માં દરિયા કિનારે અનેક જીવ સૃષ્ટિથી ભરેલ દરિયાઈ વિસ્તાર છે અને [redacted] સમુદ્ર [redacted] લાંબો છે\nઅને હાલના ઘણો ખરો વિસ્તાર પ્રદુષણ મુક્ત છે. અને આ પ્રોજેક્ટ અંતર્ગત જો દરજોજનું [redacted] [redacted] કેમિકલ વાળું છોડવામાં\nઆવશે તો દરિયાઈ જીવનું નિકંદન નીકળી જશે અને સાથે [redacted] માં રહેવા વાળા લોકોના જીવ પણ જીવા જેવા નહી રહે. જેમના\nહિસાબે [redacted] લોકોને રોજગારી ઉપરાંત સારું જીવન જીવવી શકવાની શક્યતાઓ નહીવત થાય છે અને અને લોકો ભોગ બનશે. [redacted] ના કારખાના નું\nપાણી દરિયામાં નાખીને ધંધો રોજગાર નહી છીનવાઈ તો સરકારશ્રીની ભૂલ છે કેમ દરિયાઈ વ્યાપાર સાથે [redacted] નો [redacted] મુખ્ય તેમજ\nલાખો લોકો [redacted] જુલ્લામાં દરિયાઈ ખેત ઉત્પાદન સાથે સંકળાયેલ છે અને આવનાર સમયમાં જો પ્રોજેક્ટ અંતર્ગત જો દરિયામાં દુષિત પાણી\nઠલવવાનું થશે તો આ તમામના રોજગાર-ની સાથે સુંદરતા પણ [redacted] ની છીનવાઈ જશે. વિશ્વ પ્રસિદ્ધ માછલીઓ, પરવાળા અને તથા અહીંયા\nના કિનારાઓ અદભુત સૌદર્ય ધરાવે છે. ગ્રીન સી ટર્ટલ, ડોલ્ફિન, વેલ શાર્ક, જેવી અલૌકિક જીવ સૃષ્ટિ જે વિશ્વ માં વિલુપ્તતા ને\nઆરે છે તે [redacted] ના દરિયા માં ખુબ આસાની થી અને વધુ પ્રમાણ માં જોવા મળે છે જે [redacted] વાસીઓ\nમાટે ગર્વ ની બાબત છે. આ કેમિકલ થી આ તમામ જીવ સૃષ્ટિ નો નાશ થઈ જશે... વિદેશ થી આવતા અલભ્ય\nપ્રજાતિ ના અસંખ્ય પક્ષીઓ જે આ વિસ્તાર ની મુલાકાત લે છે, [redacted] થી [redacted] મહિના અહીંયા વસવાટ કરે છે તે\nપણ આવતા બંધ થઈ જશે. આ પાણી [redacted] ના તળ ની અંદર આવી જશે તો [redacted] ના નાગરિકો ની સ્વાસ્થ્ય\nની સુરક્ષા માટે મોટો ખતરો બની જશે. [redacted] ની જેમ ડાઇંગ ઉદ્યોગ પર નભે છે તેમ [redacted] ખિલ્લે જમીન કૃષિ\nઅને દરિયાઈ કૃષિ એમ બે પ્રકારની કૃષિ પર નભતો પ્રદેશ છે આ કેમિકલ વાળા પાણી થી આસપાસ ની દરેક કળદ્રુપ\nજમીન બંજર માં ફેરવાય જશે અને દરિયાઈ જીવ પણ	Gujarati
DHLTH/E/2020/[redacted]	Health & Family Welfare >> Public Health >> Disease Outbreak >> Other ---------- कोविड 19 (कोरोना वायरस) में एक\n[redacted] जिला [redacted] का नागरिक हूँ। आज दिनाक [redacted] को दो व्यक्ती पाउडर वाली दवा मुफ्त में\nबाट रहे थे उन्होंने किसी भी प्रकार का मास्क भी नहीं पहना हुआ था। वह दोनों काफी नजदीक आकर बात\nकर रहे थे उनसे संक्रमण होने का खतरा हो सकता है। वह खुद को [redacted] का बता रहे थे जोकि [redacted] से लगभग [redacted] की दूरी पर है। जबकि लॉक डाउन चलते एक शहर से दूसरे शहर जाना ठीक\nनहीं है।	Hindi
DARPG/E/2017/[redacted]	This is complaint on behalf of residents of [redacted], [redacted], [redacted] regarding dilapidated condition of internal colony road\nin [redacted] road which leads to [redacted] Maal, [redacted] which is causing problems and difficulties for the\nresidents of the localities. This road is about [redacted](approx). The road has become totally damaged due to potholes. The\nroad needs immediate attention of the authorities concerned as residents of the area have been facing inconvenience.This road leading to\nMarket "[redacted] Maal is used by a large number of people of the area. After seasonal rains, the accumulation of\nwater causing several diseases in the area as it becomes very difficult to use this road due to damaged road.\nThe authorities concerned need to devise plan for keeping the roads and streets clean and neat to avoid any kind\nof health issue for the residents.	English
DHLTH/E/2020/[redacted]	Health & Family Welfare >> Public Health >> Non-communicable Diseases ---------- विषय :- राष्ट्रीय स्वास्थ्य मिशन अंतर्गत कार्यरत संविदा प्रबन्धकीय\nअधिकारीयों तथा कर्मचारियों के नियमितीकरण विषयक ।	Hindi

Figure 4.1: Semantically Similar Multilingual Grievances. Tokens that could reveal personal identifiable information (PII) have been redacted.

2. Query: bank collected extra amount and frauds

reg_no	subject_content	Language
CBODT/E/2017/[redacted]	ASSESSMENT YEAR [redacted] :- Out standing tax remand of [redacted] for assessment year [redacted] is reflecting on site, for your kind\ninformation this demand was reduced to [redacted] by the assessing officer and payable amount has been deposited, proof of demand\nis enclosed for your kind verification, it was intimated to the department for filling of grievance date [redacted] but demand\nnor removed till date. secondly Income Tax Refund advice for assessment year [redacted] was produced the assessing Officer for issuance\nof new refund advice with correct bank particulars as on [redacted] in circle [redacted] it is pending till date and\nno any intimation or refund check has been released, another side demand of [redacted] for assessment year [redacted] is appearing\non site. Therefor it is requested to remove the the both demand and released fresh refund advice with new bank\nparticulars and oblige.	English
DEABD/E/2020/[redacted]	वित्तीय सेवाएं विभाग (बैंकिंग प्रभाग) >> Misbehaviour/Harrassment/Corruption by Bank Staff बैंक / वित्तीय संस्थान : Other / Private Sector Bank.\nशाखा / बैंक और शाखा का नाम : [redacted] ---------- शिकायतकर्ता द्वारा बताया गया कि ग्राम [redacted] से आवेदक का नाम [redacted] [redacted] बाताया गया है आवेदक के खाते में बहुत ज्यादा पेसे काट रहे\nहै इस कारण आवेदक अपना खता बंद करवान चाह रहा है बैंक से बैंक से खाता बंद नही कीया जा\nरहा है जिससे समस्या हो रही है बैंक बैंक का नाम [redacted] का नाम कृपया समस्या का\nजल्द से जल्द निराकरण किया जाए .	Hindi
DEABD/E/2021/[redacted]	वित्तीय सेवाएं विभाग (बैंकिंग प्रभाग) >> Misbehaviour/Harrassment/Corruption by Bank Staff बैंक / वित्तीय संस्थान : Other / Private Sector Bank.\nशाखा / बैंक और शाखा का नाम : [redacted] जिला [redacted] ---------- स्पष्टीकरण किए जाने के संबंध	Hindi
DEABD/E/2020/[redacted]	वित्तीय सेवाएं विभाग (बैंकिंग प्रभाग) >> Misbehaviour/Harrassment/Corruption by Bank Staff बैंक / वित्तीय संस्थान : Other / Private Sector Bank.\nशाखा / बैंक और शाखा का नाम : [redacted] ---------- शिकायतकर्ता द्वारा बताया गया है की आवेदक के\nसाथ शाखा परिसर में अभद्र भाषा का प्रयोग किया गया है।	Hindi
CBODT/E/2018/[redacted]	I have a Senior Citizen saving scheme [redacted] account No. [redacted] at [redacted], [redacted] branch, [redacted] [redacted] IFSC Code [redacted] I have deposited Rs [redacted] and received a sum of Rs [redacted] every Quarter. In the\nquarter ending [redacted] the bank credit a interest of Rs [redacted] which is less by Rs. [redacted] and\ndeducted tax [redacted] and [redacted] which is not reflected in [redacted] of income tax. I approached the bank manager\nfor the said anomaly and promised to rectify the same by [redacted] I regret to inform you that\nthe anomaly is not rectified. That means the tax deducted is not deposited to income tax department and no refund\nclaim could be submitted. It therefore humbly requested a instruction may please be given to the bank to credit the\nwrong deducted TDS.	English
CBODT/E/2018/[redacted]	RESPECTED SIR WE HAVE APPLIED FOR PAN UNDER CATEGORY ASSOCIATION OF PERSONS (ACK [redacted] dated [redacted] WHEN WE CHECK\nSTATUS, WE FOUND THE QUERY THAT Proof of identity not provided/ not valid. WE AGAIN SUBMIT THE PROOF OF IDENTITY,\nTHEN WE TRIED TO CALL TIN FACILITY CENTER, THEY PROVIDED US MANY OF NUMBERS LIKE [redacted], [redacted] BUT NONE OF NUMBER IS PROVIDED ANY SOLUTION. THEY ARE NOT ATTENDING CALL, IF\nATTENDING THEN TRANSFER TO OTHERS, WE ARE FOLLOWING LAST [redacted] TO GET STATUS OF PAN BUT WE HAVE NOT\nFIND ANY SOLUTIONS. IT LOOK LIKE ONLY HARASSMENT FROM LAST [redacted], ALMOST [redacted] PASSED, YOU ARE REQUESTED TO\nISSUE PAN NUMBER OR INFORM TO US FOR QUERY AS EARLIEST AS. PLEASE TAKE NECESSARY ACTION WE REQUIRED URGENT PAN\nTO OPEN BANK ACCOUNT AND OTHER FORMALITIES	English

Figure 4.2: Semantically Similar Multilingual Grievances. Tokens that could reveal personal identifiable information (PII) have been redacted.

4.8 Conclusion and Future Work

In conclusion, this chapter presents a pipeline for building a similarity search system for subject-content matching in a multilingual setting. The pipeline includes various steps such as text cleaning, language detection, machine translation, and sentence embedding. The system uses a pre-trained multilingual model for machine translation and a pre-trained sentence embedding model for computing sentence embeddings. The hnswlib package is used to build the index for similarity search.

Future work can be done in several directions. Firstly, the system can be extended to support more languages and their corresponding Unicode ranges. Secondly, more advanced techniques can be employed for text cleaning and preprocessing to improve the accuracy of the system. Thirdly, the system can be evaluated and compared with other existing systems for similarity search in a multilingual setting. Lastly, the system can be integrated into an application for practical use in the industry.

Bibliography

[1] facebook/m2m100_418M. https://huggingface.co/facebook/m2m100_418M, 2021.

[2] l3cube-pune/hing-bert. https://huggingface.co/l3cube-pune/hing-bert, 2021.

[3] l3cube-pune/hing-mbert. https://huggingface.co/l3cube-pune/hing-mbert, 2021.

[4] salesken/similarity-eng-hin_latin. https://huggingface.co/salesken/similarity-eng-hin_latin, 2021.

[5] Faisal Ahmed and Abu Sayed Md Choudhury. A framework for citizen engagement in public service delivery. *International Journal of Public Administration in the Digital Age*, 5(1):41–57, 2018.

[6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1311, 2018.

[7] Debswapna Chakrabarti and Pabitra Mitra. A dynamic programming approach to segmentation of roman texts. *Journal of King Saud University-Computer and Information Sciences*, 29(2):171–178, 2017.

[8] Shikha Choudhary, Rohit Singh, Arpan Dutta, and Gurmukh Singh. A machine learning approach to address public grievances in india. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 301–305, 2018.

[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Multilingual word embeddings from wikipedia. *arXiv preprint arXiv:1801.06126*, 2018.

[10] Deepali Joshi, Nitesh Yadav, Atul Dixit, and Gaurav Singh. Design and implementation of an online public grievance redressal system. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(6):697–700, 2015.

[11] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2008.

[12] Yinan Liu, Mingdong Wang, Guangxiang Wang, Yong Chen, and Jun Li. Efficient k-nearest neighbor search using hnsw. In *Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data*, pages 649–664. ACM, 2018.

[13] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *Information Systems*, 75:41–56, 2018.

[14] Yu A Malkov and D A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. In *Advances in Neural Information Processing Systems*, pages 3847–3857, 2018.

[15] James Martin. Spelling correction and the edit distance algorithm. *Linguistic Data Consortium, University of Pennsylvania*, 4(2):1–10, 1999.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] Md Saifur Rahman, Md Habibul Islam, and Tasnim Mahmud. Sentiment analysis of citizen complaints for government services improvement. *Journal of Information Technology and Economic Development*, 11(2):14–29, 2020.

[18] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt Publishing, 2017.

[19] Partha Saha, Sourish Bhattacharya, Anirban Das, and Anirban Basu. Improving public grievance redressal system using social media analytics. *2019 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 1321–1326, 2019.

[20] Vikas Singh, Prakhar Gupta, Harsh Shrivastava, Ravi Kiran Sarvadevabhatla, and Manish Shrivastava. Hinglish-top: A large scale code-switched semantic parsing dataset, 2022.

[21] Esko Ukkonen. Fast approximate string matching with large edit distances. *Algorithmica*, 6(3):251–260, 1985.

[22] George Kingsley Zipf. A study of frequency distributions of english letters and some words. *Journal of the American Statistical Association*, 30(201):28–33, 1935.