# Poetry Identification and Recommendation System

Anup Jha
MT21163
anup21163@iiitd.ac.in

Debarshi Dasgupta
MT21167
debarshi21167@iiitd.ac.in

Archit Arora
MT21164
archit21164@iiitd.ac.in

Manish Jethmalani
MT21169
manish21169@iiitd.ac.in

Robin Kumar
2019093
robin19093@iiitd.ac.in

## Abstract

*Oftentimes, we hear a line or a verse penned by a poet but we are not able to ascertain the following lines, the name of the poem or poet. In addition, many poetry and ghazals written by poets of Indian origin are available on internet mostly in* Hinglish *language. Hence, to entertain the inquisitive minds of poetry aesthetes, we propose a web application titled as "**Poetry Identification and Recommendation System**" which is an exemplar of* Cross Language Information Retrieval System. *The novelty of our work is that it tackles with the spelling correction and mapping of Hindi and Urdu words written using Latin alphabets to the closest correct word and finally it not only retrieves the desired poetry document but also recommends some other poems based on the search results from the input query. We achieve this using novel algorithm proposed in this work, called as **HMuM**(*Hinglish Mapping using Metaphone*) . The evaluation section depicts the impressive qualitative and quantitative results of our proposed project.*

## 1. Introduction

### 1.1. Motivation

India being a place which has a significant Hindi speaking population,the poetries and other hindi creations are liked by a large pool of people across the nation. Presently, thousands of hindi poems, shayaris and other category of beatifully written lines are available on the web, but most of us are exposed to a limited part of it.

People often come to know about any poetry work through books, various stages delivering such content and social media. We enjoy it momentarily and tend to forget it so recalling a general or situation specific poem when we wish to, becomes a tough job.

Therefore, a Poetry Identification and Recommendation System can be helpful to filter and identify a poem and recommend few similar ones according to the need of users.

### 1.2. Problem Statement

The "Poetry Identification and Recommendation System" will retrieve information from a data-set which is scrapped from web and stored in a database based on the query provided by the user and will show the list of poetry consisting of the query entered, as well as it will tell the category of poetry along with some suitable recommendations.

The input to our system will be query in "hinglish" i.e. we will be feeding the english typed word of hindi queries such as "zindagi" for devanagari script of *Zindagi*. This is to be noted that the input does not have to be devanagari script instead the transliterated hindi words are to be given to the system. further output will be the complete poetry along with details like its writer, category, Recommendations etc.

To develop such system, we need an enormous amount of data which we have scrapped from Rekhta - A public platform for urdu and hindi poetry. A hinglish poetry dataset will be collected and used for the query entered. One of the major challenges in this project is to handle the noisy query, which means either the query has a typing error or after transliteration there could be more number of ways for writing a same Hindi word; for example the hinglish word for *You/Your's* could be either "aapke", "apke" or "aapkey".

## 2. Literature Review

[1] created a dataset of lyrics of sev- eral Hindi songs using a web crawler developed by them which comprised of many html tags and irrelevant characters. They performed various pre-processing steps like removal of html tags , dupli-cate characters,stopwords and proposed an unsu-pervised stemming algorithm to handle noisy data. They provided a baseline system for identifying lyrics using fuzzy c-means and improved by means of DOC2Vec and SOFM based systems.However ranking of the retrieved

lyrics was not considered while evaluating the system and this may be con-sidered as a drawback.

[2] proposed an approach in-volving retrieval of lyrics from lines of sung queries.They extracted phoneme posteriorgrams and then mapped it to a symbolic sequence and compared it using a modified Levenshtein distance with that present in the lyrics database .

[3] proposed a text input base music information retrieval system to retrieve music from the database. Each input sentence is broken down into three-four phrases and compared with lyrics in the database. The retrieval procedures were accompanied by key- word spotting system and HMM as the sequential steps.

[4] developed a mu- sic application that lists songs based on the input given in the form of text or audio. Information re- trieval technique like Vector Space Model was used.They started with preprocessing of lyrics files and search query. The term matrix for the lyrics file was calculated. Term frequency and Inverse document frequency were obtained for lyrics file and query. For checking similarity, cosine similarity between each lyrics file and query was calculated. In the end, the lyrics files were sorted based on cosine similarity values and mapped with the song files.

[5] built a search engine that can be queried through natural language. The process of deriving descriptions was automated using methods from web retrieval and music information retrieval.For each track, they retrieved a set of web pages via Google. This helped in combining information about the context of music with information about the content. The extracted text-based information is complemented by audio-based similarity, which improved the results of the retrieval by reducing the dimensionality of the feature space.

[6] proposed a procedure for recommendation of music using lyric network based on extracted keywords. Keywords representing music were extracted from its lyric by combining TF-IDF method and principle of discriminant analysis. Many experiments were carried out for analysing the lyric network judging its effect on recommendation.

[7] used TF-IDF technique as a baseline for text representation and then applied PCA to significantly reduce the dimension. They also used word- level embeddings methods to maintain proximity of words. A comparison was drawn between lyrics based methods and audio based methods for music recommendation with lyrics based methods slightly outperforming the latter.

[8] used Artificial Neural Networks and KNN regression as the algorithms for prediction of similarity score between songs. The keywords in each song are used to assign a lyrics score using BOW and finally a MYSQL database was used to aggregate the data. The decreasing graph of loss function with increasing number of epochs was compared for both ANN and KNN.

[9] displayed the impact of spelling correction on an information retrieval task by demonstrating the same by using a Pos representation of the context, along with a traditional language. This resulted in improvements compared to other versions. An observation was made that results were always improved when the addition of a contextual classifier like Word, Pos, or Posword was done. It was also shown that spelling correction with pre-processing using the left context was outperformed by the same when using both left and right contexts.

[10] implemented the TF-IDF algorithm on a number of documents and the strengths and weaknesses of the algorithm were compared. A survey was prepared on how various researchers improved/proposed solutions for the existing problems by using various improved versions of TF-IDF. It was observed that the algorithm can be applied in cross-language with the help of statistical translation. It was also stated that combining TF-IDF with other algorithms like Naïve-Bayes proved to give better results.

## 3. Baseline results (system/prototype)

Github Link

So far we have implemented 3 methods namely Posting Lists, TF-IDF, and BM25 out of which TF-IDF and BM25 scores were used to get a list of relevant documents sorted according to their ranks. Whereas the posting list returns the list of documents containing the query in the exact sequential order.

To take care of the accent and writing style issues from the raw data we have performed the pre-processing for both our data-set and query entered. Pre-processing steps include accent removal, punctuation removal, conversion to lowercase etc.

Following results were obtained for the query ”lakh talwaren badhi aati hon gardan ki taraf sar jhukana nahin aata to jhukaen kaise” :

1. Top documents from BM25 methods are:

```
[('apne chehre se jo zaahir hai chhupaaen kaise wa_waseem barelvi.txt',
  75.22713524444431),
 ('ab jo koii puuchhe bhii to us se kyaa sharh e haalaat karen fai_faiz ahmad faiz.txt',
  16.6437642482313),
 ('vo mujh ko kyaa bataanaa chaahtaa hai wa_waseem barelvi.txt',
  15.024245292489168),
 ('bulaatii hai magar jaane kaa naiin _rahat indori.txt', 14.592487563330831),
 ('jo dil pe guzartii hai vo samjhaa nahiin sakte shakee_shakeel badayuni 3.txt',
  14.034459484790927)]
```

2. Top documents from TF-IDF methods are:

```
Top 5 relevant documents along with their smilarity scores is as following:
apne chehre se jo zaahir hai chhupaaen kaise wa_waseem barelvi.txt >> 1.0000000000000002
vo mujh ko kyaa bataanaa chaahtaa hai wa_waseem barelvi.txt >> 0.43593030211396583
taraash kar mire baazuu udaan chhod gayaa pa_parveen shakir.txt >> 0.37427612292857315
vo sivaa yaad aae bhulaane ke baad khumar_khumar barabankavi.txt >> 0.3706560096961068
khirman e jaan ke liye khud hii sharar ho gae ham p_pirzada qasim.txt >> 0.36911371236419116
```

3. Top documents from Positional Index methods are:

```
The number Poems consisting the entered query is :  1
Entered query belong to the Poem:  ['apne chehre se jo zaahir hai chhupaaen kaise wa_waseem barelvi.txt']
```

The baselines models were efficient in retrieving the documents but they have some shortcomings too. All the models work efficiently until there is no typing error in the query and the spelling of words in the query are consistent with that of corpus. Hence, the baseline models are not robust enough to handle the erroneous inputs and might throw a Key-Error for the same.

## 4. Methodology

The first task designated for the progress of the project was to implement the baseline models. Out of the three methods implemented as baseline models, TF-IDF methodology of generating vectors for each document is employed. Further, for implementation of the Vector Space Model, treating each document as well as query as a vector cosine similarity score for each document is calculated respective to each query. Thereafter, each document is ranked on the basis of the scores obtained.

The second challenge was to deal with the words in the query which are either phonetically similar or have some potential typing error(consider the spelling of the same word from the corpus to be the standard). The phonetically similar pair could be (”karib”, ”qareeb”), (”hain”,”hai”), (”bulati”, ”bulatey”) etc. Similarly, the spelling error could occur in form of either repetition of a letter or producing a misspelt word by user(human error). To deal with both the cases, first a *metaphone map* was created which contains keys as *metaphone code* and values as a list of words corresponding to that Metaphone code. Further, if the word from the query exist in the corpus of words then it is mapped to itself. Moreover, if the word is not present in the corpus but the Metaphone code for the word is present in the *Metaphone Map*, then that word is mapped to closest word in the list obtained from *Metaphone Map* which has least edit distance. Lastly, if the Metaphone code is also not present in the dictionary then a list is prepared by concatenating all the lists corresponding to the keys which has least edit distance with respect to the code of the word. Now, to obtain the closest word in the this case, the final amalgamated list is used. As a result of the mentioned algorithm, all the words get mapped to some word from the corpus. We state this algorithm of spell correction as **HMuM**(*Hinglish Mapping using Metaphone*) .After, successfully dealing with the input query, the processed query is ready to be utilized for the first task as discussed previously.

To generate recommendations, the ’candidate poets’, who’s poetry matches with query are selected. Now, based on the ranking of the ’candidate poets’, their random poetry works are extracted and are recommended to the user. Finally, the webapp, title as *Poetry Identification And Recommendation System*, was hosted using a popular framework known as ”Flask”.

## 5. Contribution

1. A dataset has been created by scrapping the website Rekhta. The dataset is stored as a collection of text files such that each file contains a *Ghazal* which is penned by some well known poet. The dataset can be found as ”Poems” folder in the github repository.

2. The most important and novel contribution of the project is through the proposition of **HMuM**(*Hinglish Mapping using Metaphone*) algorithm as a part of preprocessing step.**HMuM** is not only capable of handling the edge cases of handling the phonetically similar words but also is able to map the unknown words to the closest word from the corpus using *Metaphone map*.

## 6. Evaluation

The project work is an exemplar of application of concepts to improve upon the engineering problem of **Cross Language Information Retrieval**. After the mapping of each word in the erroneous query, the retrieved documents have similar ranking as that of the case in which query was perfect. The results for three examples are as following:

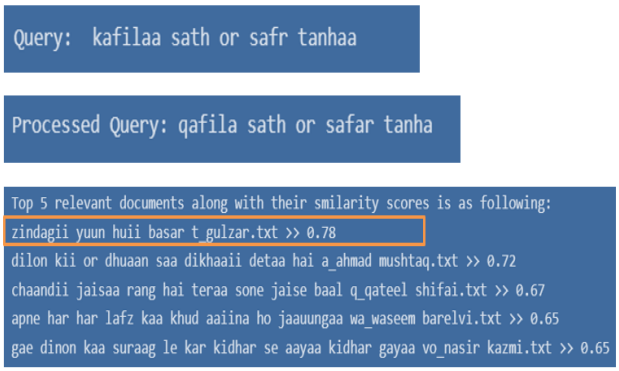
1. Test Case 1:

```
Query:  bulatiiiin haii magur jaaney ka nai
```

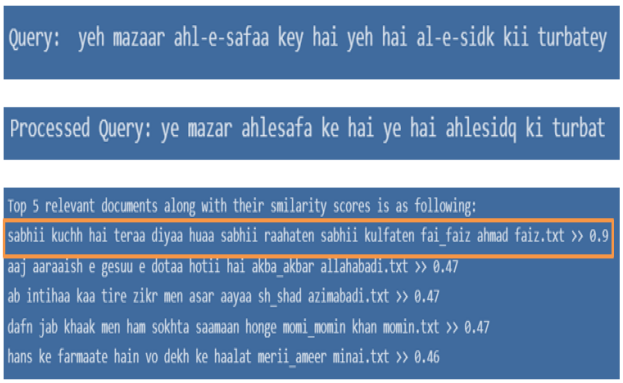
```
Processed Query: bulati hai magar jaane ka nai
```

```
Top 5 relevant documents along with their smilarity scores is as following:
bulaatii hai magar jaane kaa naiin _rahat indori.txt >> 0.88
tujh se bichhde hain to ab kis se milaatii hai ha_shahryar.txt >> 0.85
dhuum se sunte hain ab kii saal aatii hai bahaar sauda m_sauda mohammad rafi.txt >> 0.8
zindagii jab bhii tirii bazm men laatii hai ha_shahryar.txt >> 0.8
aap kii yaad aatii rahii raat bhar fai_faiz ahmad faiz.txt >> 0.78
```

2. Test Case 2:




3. Test Case 3:




From the above test cases it could be stated that the application is retrieving the documents efficiently as the ground truth document(highlighted by orange rectangle) is at first place for all the three cases.

## 7. Conclusion

The implemented application is efficient enough to handle the erroneous queries and it produces results consistently. From the test cases stated in the evaluation section it could be observed that how an *Hinglish* query is being able to be mapped to a correctly spelled query(according to corpus), thus being a perfect example of implementation of a **Cross Language Translation** and **Cross Language Information Retrieval** system. In addition, the *Poetry Identification And Recommendation System* is developed as an webapp which is able to cater to the inquisitive mindset of poetry appriciators.

## References

[1] B. G. Patra, D. Das, and S. Bandyopadhyay, "Retrieving Similar Lyrics for Music Recommendation System," p. 8. 1

[2] A. M. Kruspe and M. Goto, "Retrieval of Song Lyrics from Sung Queries," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Calgary, AB, Canada), pp. 111–115, IEEE, Apr. 2018. 2

[3] T. Wang, D.-J. Kim, K.-S. Hong, and J.-S. Youn, "Music Information Retrieval System Using Lyrics and Melody Information," p. 4. 2

[4] P. Pulicherla, A. Prakash, and J. S. Bhanu, "Retrieving Songs By Lyrics Query Using Information Retrieval," vol. 8, no. 6, p. 3. 2

[5] P. Knees, T. Pohle, M. Schedl, and G. Widmer, "A music search engine built upon audio-based and web-based similarity measures," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, (Amsterdam, The Netherlands), p. 447, ACM Press, 2007. 2

[6] K. Nakamura, T. Fujisawa, and T. Kyoudou, "Music recommendation system using lyric network," in *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, (Nagoya), pp. 1–2, IEEE, Oct. 2017. 2

[7] M. Vystrčilová and L. Peška, "Lyrics or Audio for Music Recommendation?," in *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, (Biarritz France), pp. 190–194, ACM, June 2020. 2

[8] "Music Recommendation System," *International Journal of Engineering Research*, vol. 8, no. 07, p. 2. 2

[9] P. Ruch, "Information Retrieval and Spelling Correction: an Inquiry into Lexical Disambiguation"," p. 5. 2

[10] S. Qaiser, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications*, vol. 181, p. 5. 2