CS-565 INTELLIGENT SYSTEMS AND INTERFACES

PROJECT PROPOSAL

ABHISHEK JAISWAL (170101002)
HARDIK KATYAL (170101026)
MANAN GUPTA (170101035)
MAYANK WADHWANI (170101038)

PROBLEM STATEMENT

Poetic artistry is a child of linguistic arts which, much like its siblings, visual and performing arts is heavily influenced by its creator and his ideology. Even in the greatest of masterpieces, we see the influence of the era in which it was written materializing itself in subtle forms of expressions and vocabulary usages.

The works of Kabir, one of the most well-known poets of Hindi, can be seen to be influenced by the Bhakti Movement. He wrote most of his poems in vernacular Hindi frequently borrowing words from various dialects like Braj encompassing topics like devotion, discipline, and mysticism.

All these intricate features including the use of beguiling metaphors, rhyme schemes, and the innate preference of an individual to a specific style of penmanship and vocabulary create an elaborate interplay almost akin to a fingerprint. These features can therefore be used for predicting the author and era of previously unseen poems. This task has been effectuated for the English language but Indic languages like Hindi have largely remained untouched. We therefore propose to implement a model that handles this task for Hindi poems.

MAJOR CHALLENGES

There are various challenges that we may encounter during the course of this project, the most prominent one being of trying to fathom the convoluted features from the few lines of text as input.

It would also be a challenge to get a complete set of vocabulary and word embedding because of the extensive use of vernacular dialects prevalent across different eras and because the language Hindi in general has evolved so much across the scores of centuries.

Creating a corpus will pose yet another challenge as there is no available corpus that we can use at this time.

BRIEF REVIEW OF EXISTING MODELS

As mentioned above, since not much work has been done in Hindi Language, we present below the work that has been done in different languages.

In his paper, Vaibhav Kesarvani has made an attempt to perform automatic poetry classification for English poems using NLP. He has tried to prove some interesting and important hypothesis such as automatic poetry diction is possible by using word embedding as features. He has used several other features and quantified them to use them as parameters to train a CNN based model. These quantifiers may extend to include several features like verbal density of the poem, etc. There also exists various other works (for ex.

POEMAGE) that tries to classify a poem based on sonic elements, but we restrict ourselves to text only. This will help us to make use of various NLP and ML models. There are also tools like SPARSER which aims to study poetry by the use of NLP tools like tokenization, sentence splitters, NER (Named Entity Recognition tools) and taggers. However since the corpus (consisting of around 500 poems) is not very large, the efficiency of the SPARSER tools is decent. We thus can conclude that automatic poem classification is possible. There are other tools like VerseVis which helps us to visualize rhyme and meter in a poem in a visually colour coded manner.

More works includes designing of an emotion classification and poet identification model that uses SVM classifier to achieve 92.3% accuracy. It uses 3 stylometric features namely orthographic, syntactic & phonemic for classification. Basic blocks of SVM classifier are alliteration, reduplication, rhyme & document statistics.

Also, text-mining methods have been developed which when applied to the poems helps in recognizing the author of the text. In pre-processing step, all the words are converted from uppercase into lowercase, punctuation marks and digits are deleted, tokenization is done according to the white space character, stemming type of all the words is identified and stop words are deleted. TF.IDF method have been used to apply term weighting.

These works helps us to believe that automatic rhyme analysis and rhyme quantification is possible and various features of a poem can be used for the same.

PROPOSED DIRECTION

We will first start by identifying major authors in different eras of Hindi literature and create a corpus containing their major works. Since no such annotated corpus currently exists, it will take sufficient effort in accomplishing this task.

Once this is done we will contemplate and argue on the best model to use for predicting the author and era. We will also be exploring any libraries we can use for word tokenization and word embedding among other tasks.

We will conclude our project by measuring the accuracy of the model by random sampling and if time permits try and extend the work to include other features of poetry like topic, rhyme scheme etc as well.

RELEVANT REFERENCES

KESARWANI, V. (2018). AUTOMATIC POETRY CLASSIFICATION USING NATURAL LANGUAGE PROCESSING. *Retrieved from https://ruor.uottawa.ca/bitstream/10393/37309/1/Kesarwani_Vaibhav_2018_thesis.pdf*

Rakshit, G. (n.d.). Automated Analysis of Bangla Poetry for Classification and Poet Identification. *Retrieved from http://cdn.iiit.ac.in/cdn/ltrc.iiit.ac.in/icon2015/icon2015_proceedings/PDF/12_rp.pdf*

Sennrich, R. (n.d.). Neural Machine Translation of Rare Words with Subword Units. *Retrieved from https://www.aclweb.org/anthology/P16-1162.pdf*