INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

SUKHAN: Corpus of Hindi Shayaris annotated with Sentiment Polarity Information

ABSTRACTS

Shayari is a form of poetry mainly popular in the Indian subcontinent, in which the poet expresses his emotions and feelings in a very poetic manner. It is one of the best ways to express our thoughts and opinions. Therefore, it is of prime importance to have an annotated corpus of Hindi shayaris for the task of sentiment analysis. In our work, we introduce SUKHAN, a dataset consisting of Hindi shayaris along with sentiment polarity labels. To the best of our knowledge, this is the first corpus of Hindi shayaris annotated with sentiment polarity information. This corpus contains a total of 733 Hindi shayaris of various genres. Also, this dataset is of utmost value as all the annotation is done manually by five annotators and this makes it a very rich dataset for training purposes. This annotated corpus is also used to build baseline sentiment classification models using machine learning techniques

MOTIVATION

The task of sentiment analysis becomes challenging for languages having annotated corpus only in some limited domains. One such language is Hindi and shayari is one of its domains which has no annotated dataset for the task of sentiment analysis. Shayari is a very rich tradition in South Asia. It has generally 2 to 4 lines which have some kind of deep meaning in them. It is mainly written in languages like Hindi, Urdu, Bangla, Nepali and Punjabi. Whether you are sad, alone, happy or in love, you can use shayari to express your feelings and thoughts. That's why, it is very important to have an annotated corpus of shayaris for the task of sentiment analysis. No such annotated corpus of Hindi shayaris currently exists in literature. SUKHAN is the first corpus of Hindi shayaris with annotated sentiment polarity information existing in literature as per our knowledge. It is written in Devanagari script and hence avoids the pre-processing cost of text normalization

METHOD

- **Dataset**
  - Online websites
  - Devanagari Script
- **Annotation**
  - Done at whole level
  - Done using Russell's Circumplex model.
- **BaseLine Experiments**
  - 5 Fold Cross Validation
  - TF IDF features used
  - Word and Character n grams
  - Linear SVM performed best

Authors:- Salil Aggarwal, Abhigyan Ghosh, Radhika Mamidi

contact info: salil.aggarwal@research.iiit.ac.in

Research Center Name: LTRC