Introduction/Motivation

Sentiment analysis is the task of classifying the polarity of a given text at the document/sentence/feature level into positive or negative (and in some cases neutral) class. Most of the research in this domain has been focused on English language and work has been done at syntactic, semantic and discourse levels. Previous work on Sentiment analysis of Indian Languages has mainly focused on supervised methods though. We propose to detect the sentiment of Hindi language in devanagari text, using semi-supervised approaches.

Challenges: Hindi is morphologically rich and is a free order language as compared to English. Also, the scarcity of resources for the Hindi language brings challenges ranging from collection and generation of datasets. [1]

Related Work

Sentiment analysis for Indian Languages has primarily been focussing on using:
- Machine Translation to translate the data in English to Hindi.
- Bi-Lingual dictionary for English and Indian Languages
- Hindi WordNet expansion to exploit synonyms and antonym polarity

Recent contribution in Hindi Subjective Lexicon was done by Bakliwal et al. [2], where they created a resource for Hindi Polarity Classification. They used Hindi WordNet to retrieve synonyms and antonyms of a given word in hindi for which they knew the polarity and then assigned the similar polarity to synonyms and opposite polarity to antonyms.

Labelling emotion in Bengali blog corpus by Das et al. [3]. They manually annotated the corpora at sentence level and then used some standard techniques like Kaapa and MASI to handle the differences between the annotators.

Joshi et al., 2010 created H-SWN (Hindi-SentiWordNet) using two lexical resources namely English SentiWordNet and English-Hindi WordNet Linking. Using WordNet linking they replaced words in English SentiWordNet with equivalent Hindi words to get H-SWN.

Our Approach

Dataset: We plan to apply sentiment analysis on short (sentence/paragraph level) hindi movie and product reviews (with possible extension to unstructured data such as tweets). The reviews will be collected from websites such as jagaran, and manually/semi-automatically (based on rating) tagged. Due to the low resource availability, creating this tagged dataset will also be an important contribution of our project.
Ex of a positive movie review: ‘भाग मिल्खा भाग’ नेहरू युग के एक महत्वाकांक्षी युवक के सामने सपने की दास्तान है, जिसे राकेश ओमप्रकाश मेहरा ने बढ़ाती उस दर्शक के परिवेश और समाज के साथ प्रस्तुत किया है। यह फिल्म फिशर और युवा दर्शकों को अवश्य देखनी-दिखानी चाहिए।’

Ex of a negative movie review: ‘शॉर्टकट रोमियो’ एक ऊबाऊ फिल्म है। अमीषा पटेल, नील नितिन मुकेश, बंदी ग्रेटल और राजेश नूबापुरे चारों मुख्य अभिनेताओं ने निराश किया है।’

Technique: We plan to use semi-supervised techniques since Hindi has very little tagged data. Also, semi-supervised methods have not been explored much by earlier work. We propose to create tagged data structure for sentiments using a pre-annotated small dataset and then use machine learning models like HMM to push this tagging onto the whole corpora and then evaluate our model.

References