Introduction

Traditionally, companies have used surveys to collect public opinions and their sentiments about a product. This task becomes increasingly tedious because of factors like length of survey with repetitive questions, unclear language and a lack of relevance. This leads to negative response and incorrect data collection. But with the increase in use of social media, people post real time messages about their opinion on diverse range of topics including the daily products and services they use, government policies, film reviews, current issue, etc. Companies can use this readily available information to study the trends and to create and market their products in a better way. This requires building a tool that collects user data from social media sites, studies the data trends and presents this information in a way that is useful for companies.

The primary goal of our project is to collect data from Twitter posted in Hindi, build models to generate topics that trend on daily basis and study how the sentiment of the topic changes over a period of time. There has been considerable work done on sentiment analysis and topic modeling in resource rich languages like English, but there has been little work done in Indian languages like Hindi because of the sparseness of data available and lack of people tweeting in Hindi. With UTF-8 encoding and increase in activity on social media in Hindi, we have been able to stream live tweets in Hindi and develop models on Hindi Twitter data. The challenge in using Hindi Twitter data is that there is little or no annotated data for sentiment analysis. Our research certainly will contribute to natural language processing community enabling them to use our annotated dataset which is publicly available and also contribute to developing a tool which predicts the trends and sentiments on social media. We also use transliteration APIs to convert hinglish words to Hindi and improve the quality of the dataset.