ABSTRACT

Decades ago, people used to represent their opinion by writing it manually or by speaking at public places. These reviews are further taken as an advice for the betterment. To process this data was a complete manual task. As the usage of internet grew people started sharing their views regarding any entity through emails or social platforms. The intensification of data over the social media makes the task of deducing valuable information a bit complex. Automatic deduction of sentiments from web data is considered as a process of sentiment analysis. An algorithm devised for the same is known as sentiment analyser. Use of Sentiment analysers is at the peak for various enterprises to find the loopholes in their product or services. An optimal sentiment analyser is the one which works rationally as humans. The goal is thus to fill the research gaps associated with the effective sentiment processing.

Sentiment analysis integrates many subtasks i.e. Named Entity Extraction, Anaphora resolution, Sentiscore, Feature extraction, etc. Effective pre-processing yields better results for all the natural language processing tasks. The reason for pre-processing of the data is that people use slangs, long tail words, multilingual content and visual language such as emoticons. People use unstructured format of writing along with all the above mentioned categories these days. The process to handle slangs, misspelled words, etc. is called as normalization.

The primary aim of this study is to have effective pre-processing of the content i.e. normalization. Normalization here deals with two aspects: one is to deal with slangs and another is to deal with emoticons. In this study, a technique is used where each emoticon is mapped corresponding to its meaning for generating Sentiscore, instead of just adding or subtracting one for positive and negative smiley respectively. Slangs are also handled effectively by using cross word dictionary and corpus. It is aimed to get better results for pre-processing.

This thesis also puts light on how to deal with multilingualism. These days internet provides the facility to people for writing in any of the desired language or mix of different languages. It makes the task of sentiment analysis more complex. The
betterment of sentiment analyser is based on processing this data regardless of the language in which it is written. Multilingualism also comes in the form of macaronic content. In this thesis, a complete analysis of macaronic content is discussed with the proposed technique.

The next objective of the thesis is to present some new results of investigations, demonstrating an application of Temposentiscore for problems related to categorization of reviews in web. Deal with obsolete reviews is the next concern of this study. These reviews may result in biased sentiment analysis which may or may not present the current scenario. To remove this limitation, we are trying to implement temporal sentiment analysis of reviews by providing more weightage to latest reviews. Further, sentiscore is redefined in terms of temposentiscore. Finally, the study mainly emphasizes the need to fulfil the research gap.