Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 554 – 563

Procedia Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

Context Specific Lexicon for Hindi Reviews

Deepali Mishra[a], Manju Venugopalan[a] and Deepa Gupta[b,*]

[a]*Department of Computer Science, Amrita School of Engineering, Bangalore, Amrita Vishwa Vidyapeetham, Amrita University, India, deepalitiwari22@gmail.com*
[a]*Department of Computer Science , Amrita School of Engineering, Bangalore, Amrita Vishwa Vidyapeetham, Amrita University, India, v_manju@blr.amrita.edu*
[b]*Department of Mathematics, Amrita School of Engineering, Bangalore, Amrita Vishwa Vidyapeetham, Amrita University, India, g_deepa@blr.amrita.edu*

Abstract

In the era of social networking, immense amount of posts, comments and tweets generated every second are increasing the size of social database .The analysis of this voluminous data is necessary for exploring the orientation of people's opinion about a particular entity. Most of the online data are in English language, but due to increase in technology and improved awareness of people, the online data available in Indian languages are gradually increasing. Sentiment analysis of English language alone is not sufficient to know the inclination of people towards an entity, other Indian language sentiment analysis is a must, their contribution is also important for us. The available sentiment classification lexicon resources like Hindi SentiWordNet are generic in nature and hence results in average sentiment classification accuracy due to contextual dependency. To improve the sentiment classification accuracy, we present an improvised lexicon resource for Hindi language for Hotel and Movie domains. The improvised polarity lexicon has been built reflecting context sensitivity and to increase coverage it has been expanded used synonyms based approach. The built polarity lexicon resource showcases an improvement in accuracy of 42% and 78% in Movie and Hotel domain, respectively, compared to the existing Hindi SentiWordNet lexicon resource.

Peer-review under responsibility of the Organizing Committee of ICACC 2016

Keywords: Sentiment analysis ; lexicon ; HSWN ; LR ; LRE etc .

* Deepa Gupta. Tel.:+919916921850.
E-mail address: g_deepa@blr.amrita.edu

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.283

1. Introduction

The current decade has been witnessing an exponential increase in the number of users and web content. This voluminous data are used by people to get an idea in decision making about any entity. For example before travelling to any unknown place, previously we would prefer talking to those who have visited that place, but now due to online available data in the form of reviews, we go by the reviews for a decision making. These available text data need to be analyzed, and hence the opinion orientation identified which is termed as opinion mining or sentiment classification. Almost two decades of work has been contributed to extracting sentiment from English the broader categories being sentiment classification, lexicon resource creation etc. but minimal work have happened on Indian languages. The increase in the volume of Indian language data available online has elevated the importance of exploring sentiment in Indian Languages.

With the advent of technology where many social networking sites like Twitter, Facebook etc. providing provisions to express in a handful of Indian Languages, newspapers, blogs etc. providing provisions for native expressions have led to more Indian Language content available online. Even though English is an International Language, the sentiment extracted from English reviews alone cannot be considered to make final conclusions on an entity; other language inputs should also be considered. This creates the necessity to give some effort to sentiment analysis of Regional Languages.

The last few years have witnessed some authors showing their interest to mining in Indian languages but as mentioned earlier majority work contributions are in English. So it is obvious that more resources and tools are available for the same. Hindi is a well-known and widely spoken language in India. Web pages in Hindi language have increased on a rapid pace. There are many websites which provide information in Hindi owned by various news websites providing information regarding culture, music, entertainment and other aspects of arts. The web content for Hindi language has been increasing with great speed. This emphasizes the scope for further exploration of the language. But each language puts forward challenges to be encountered in terms of its syntactic and semantic structures. Hindi is a free order language with various morphological variants, spelling variance, word sense ambiguity and contextual variances. Sentiment analysis in Hindi is less explored so there is scarcity of resources and tools. Among the existing resources the most popularly used is the Hindi SentiwordNet[1]. The classification based research works using this resource have found to exhibit average accuracy which owes to the polarity lexicons not being context sensitive. Opinion words might infer different meanings in varied domains. For example "इस सैमसंग मोबाइल की बैटरी लाइफ लंबी हैं।", "फिल्म लंबी थी।". In the first sentence the "लंबी" word in battery life context expresses a positive opinion, but in the second sentence "लंबी" word in movie context conveys a negative opinion. The polarity of the word contributed by Hindi SentiwordNet is +0.5 which is sensible for the cellphone battery context but not for the movie domain. Hence this work takes a special interest towards dealing with context specificity issue. The major contributions put forward by the proposed work are

a) Proposes an algorithm to build an improvised context sensitive polarity lexicon for a particular domain.
b) Attempts improving the lexicon coverage by the Hindi WordNet based approach

The research works attempted in Hindi Sentiment analysis have been keenly studied and the findings presented in Section 2.The Corpus details are provided in Section 3, the detailed Proposed Approach in Section 4, the Results and Analysis in Section 5 and the Conclusion and Future work in Section 6.

2. Related Works

The earliest works in Hindi Sentiment analysis can be traced back to the beginning of the current decade. Most of the works attempted classification on different domains using existing resources like Hindi SentiWordNet[1]. The work has contributed SentiWordNet for the 3 Indian languages Hindi, Bengali and Telugu by using the English SentiWordNet and the subjective word list as base resource. To build the lexicon resources for target language, the experimented approaches are machine translation or dictionary based, word net based, corpus based and online game based[2]. English SentiWordNet words are translated into target language and the same polarity score has been given to target language lexicons. To increase the lexicons in the generated target language SentiWordNet used