Can only Feel যন্ত্রনা (Jantrana: Pain)!

Sentiment Analysis for Bangla

Abstract

"The question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions."

--Marvin Minsky

NLP and AI researchers of the late 80’s start realizing that machine should have sentiment to be intelligent. Therefore in the last two decades various researchers attempted textual Sentiment Analysis (SA) for various languages. In general SA refers to the application of natural language processing, computational linguistics and text analytics to identify and extract sentimental / opinionated / emotional information from text.

SA defines an overall problem, which address multiple aspects of sub-problems. Undoubtedly it is a challenging and enigmatic research problem. In this chapter we will discuss about the various challenges and solution strategies for the Sentiment Analysis (SA) from Bangla Text.

The tittle of this chapter is inspired by the most popular Bangla Sci-Fictions writer Narayan Sanyal¹. One of the best and most popular Sci-Fictions by him is Nakshatraloker Debatatma [নক্ষত্রল঱ালের দদবতাত্মা], is based on the transformation of human race from primitive creature to civilized intelligent species controlling the whole earth. Then it deals with Jupiter exploration and a super intelligent Computer HAL. The three parted book is a inspiration of 2001: A Space Odyssey² by Arthur C. Clarke³.

In his book, Sanyal named HAL as Jantra-Na (যন্ত্র-না), in our mother tongue Bengali, it ambiguously means 'not a machine' (হত্না) as well as 'pain' (হত্না).

Overview

Sentiment Analysis (SA) from natural language text is a multifaceted and multidisciplinary AI problem, tries to narrow the communicative gap between the highly sentimental human and the sentimentally challenged computer by developing computational systems that recognize and respond to the sentimental states of the human users. Bangla (ethnonym: Bangla; exonym: Bengali) is morpho-syntactically and culturally Rich language, therefore sentiment analysis from Bangla is undoubtedly tougher, discussed in the Section 1 (যন্ত্রানুভূতি).

Sentiment is not a direct property of languages; therefore an intelligent system should need some prior knowledge to act semi-mentally. Sentiment knowledge is generally wrapped into

¹ http://en.wikipedia.org/wiki/Narayan_Sanyal
computational lexicon, technically called Sentiment Lexicon. The development process of Bengali SentiWordNet has been described in the Section 2 (যন্ত্রানুভূতি-সংকলন).

Similar to classical pattern recognition problems, Sentiment Analysis is also classified into the identification and the classification genre called sentiment / subjectivity detection and polarity classification respectively. The proposed techniques for subjectivity detection and polarity classification for Bangla are elaborated in the Section 3 (যন্ত্র-বোধদয়).

The need of the end user is the driving force behind the sentiment analysis research. The end users are not only looking for the binary (positive/negative) or multi-class sentiment classification but they are more interested in aspectual/structural sentiment analysis. Therefore only sentiment detection and classification is not enough to satisfy the need of the end user. Proper Structurization of sentiments is essential before proceeding for any further granular analysis or generation and aggregation. Structurization involves identification of various aspects of a sentiment/opinion, i.e., sentiment holder, sentiment topic and so on. The whole research attempts on Structurization are described in the Section 4 (যন্ত্রানুভূতি-পর্যবেক্ষণ).

To meet the satisfaction level of end users an intelligent sentimental/opinionated information processing system should be capable enough to present an at-a-glance presentation of aggregated information, scattered over various sources/documents. Finally, textual or visual summarization, visualization or tracking of sentiment are the striking needs from the perspective of the end user. The overall summarization-visualization-tracking research attempts are described in the Section 5 (যন্ত্র-অনুভূতি-সাংক্ষেপ).

The chapter is finally concluded with the Section 6.

1 Machine’s Sentiment (যন্ত্রানুভূতি)

There is a perpetual debate about better way of collecting intelligence either by following the functional path of biological human intelligence or generating new methodologies for completely heterogeneous mechatronics machine and redefine a completely new horizon called electronic intelligence.

"How the mind works is still a mystery. We understand the hardware, but we don’t have a clue about the operating system."

--James Watson (Nobel laureate)

2 Bangla SentiWordNet (যন্ত্রানুভূতি-সংকলন)

Sentiment knowledge acquisition in terms of sentiment lexicon is the vital pre-requisite of any sentiment analysis system. Previous studies have proposed to attach prior polarity (Esuli and Sebastiani, 2006) to each sentiment lexicon, called SentiWordNet. Following the pioneering work we have developed Bangla SentiWordNet. We proposed an automatic template based online gaming technique, called Dr. Sentiment to acquire such sentiment knowledge involving Internet Population.
3 Sentiment Detection and Classification (যন্ত্র-বোধদয়)

As like classical pattern recognition problem SA is also classified into identification and classification genre called subjectivity detection and polarity classification involves sentiment detection and sentiment classification. We adopted Genetic Algorithm for sentiment detection and Sentimantics for sentiment polarity classification.

4 Sentiment Structurization (যন্ত্রানুভূতি-পর্যালোচনা)

We propose a novel structurization scheme called: 5W (Who/কে, What/কি, When/কখন, Where/কোথায় and Why/কেন) constituent extraction technique for sentiment/opinion structurization. In details this structurization helps to understand sentiment changes of any “Who/কে” regarding “What/কি” during “When/কখন” and “Where/কোথায়” and the reasons behind or “Why/কেন”.

5 Sentiment Summarization (যন্ত্রঃতিয়-ঝুঁকি-সাংরঘ)

Aggregation of information is the necessity from the end users’ perspective but it is nearly impossible to develop consensus on the output format or how the data should be aggregated. Researchers have tried with various types of output formats like textual or visual summary or overall tracking along time dimension. Several research attempts can be found in the literature on Topic-wise and Polarity-wise summarization and on Visualization and Tracking. The key issue regarding the sentiment aggregation is “how the data should be aggregated?”. Dasgupta and Ng (Dasgupta and Ng, 2009) throw an important question: “Topic-wise, Sentiment-wise or Otherwise?” about the opinion summary generation techniques. Actually the output format varies on end users’ requirements and the domain.

To overcome this research paradox we proposed the 5W constituent based summarization-visualization-tracking system (Fig. 1), falls into every genre and is the answer for the philosophical question “Topic-Wise, Polarity-Wise or Other-Wise”.

![Diagram of 5W constituent based summarization-visualization-tracking system](image.png)
**Figure 1:** A Snapshot of the 5W Summarization-Visualization-Tracking System

**Topic-Wise:** The system facilitates users to generate sentiment summary based on any customized topic like Who, What, When, Where and Why based on any dimension or combination of dimensions as they want.

**Polarity-Wise:** The system produces an overall gnat chart that can be treated as the overall polarity wise summary. An interested user can still look into the summary text to find out more details.

**Visualization and Tracking:** The system facilitates users to generate visual sentiment tracking with polarity wise graph based on any dimension or combination of dimensions as they want, i.e., “Who” are the actors and “What” are their sentiment regarding any topic, changes in sentiment during “When” and “Where” and the reasons for change in sentiment as “Why”. The final graph for tracking is generated with a timeline.

6 How far the “The Best Informed Dream of HAL or যুক্ত না”

The book “2001: A Space Odyssey” was written in 1968 and the ideological replica in Bangla by Narayan Sanyal i.e. “Nakshatraloker Debatatma [নক্ষত্রল঱ালের দেবতাত্মা]” was written in 1976, but still the HAL or “যুক্ত না” is the “The Best-Informed Dream” for AI researchers till date though we have passed approximately four decades after that science fantasy.

---

4 [http://mitpress.mit.edu/e-books/hal/chap1/one1.html](http://mitpress.mit.edu/e-books/hal/chap1/one1.html)