

# Exploiting 5W Annotations for Opinion Tracking

Amitava Das<sup>\*</sup>

Dpt. of Computer and Information Science  
Norwegian University of Science and Technology  
Sem Sælands vei 7-9, 7094 Trondheim, Norway  
amitavad@idi.ntnu.no

Björn Gambäck

Dpt. of Computer and Information Science  
Norwegian University of Science and Technology  
Sem Sælands vei 7-9, 7094 Trondheim, Norway  
gamback@idi.ntnu.no

## ABSTRACT

A system performing sentiment analysis needs some prior knowledge which can be acquired, for example, by manual annotation only or by expanding a human-annotated sentiment lexicon by automatic means. However, users are seldom interested in all sentimental aspects at once, but rather look for opinion changes of some person (“Who”) during some time period (“When”) and depending upon “What”, “Where” or “Why” something happened. The paper describes the work necessary to obtain and utilize such 5W annotations.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering;  
I.2.7 [Natural Language Processing]: Text analysis.

## Keywords

Sentiment analysis, Knowledge management, Annotation.

## 1. INTRODUCTION

The needs of the end users are the driving force behind sentiment analysis research. The end users are not only looking for binary (positive/negative) or multi-class sentiment classification, but are more interested in aspectual/structural sentiment analysis. Therefore only sentiment detection and classification is not enough to satisfy the needs of the end users. Rather, aggregation and visualization of sentiment information is a necessity from the end user’s perspective. However, a prerequisite for knowledge aggregation is having a proper and structured representation of the knowledge: sentiment structurization is a task involving the identification of various aspects of a sentiment/opinion: sentiment holder, sentiment topic, domain dependent attributes, and so on. Proper structurization of sentiments is essential before proceeding to any further granular analysis of text sentiment, or the generation and aggregation of opinionated texts.

Resource acquisition is a challenging obstacle to creating useful systems based on any type of prior knowledge and both types of acquisition techniques have their limitations. A substantial amount of work has been done on the creation of sentiment knowledge sources in several languages and domains. These efforts can be broadly categorized into two

<sup>\*</sup>A. Das is supported by an ERCIM “Alain Bensoussan” Fellowship.

types, one using only manual annotation techniques and the other based on expanding a (smaller) human-annotated sentiment resources by automatic, machine learning-based means. Automatic techniques are good for coverage expansion, but demand manual validations and are dependent on the corpus availability in the respective domain. Manual annotation techniques are more trustworthy, but generally require longer development times and a large number of annotators in order to balance the sentimentality of individual annotators—but qualified human annotators are both difficult to find and quite costly. This paper discusses 5W annotation, a way to simplify and speed up manual semantic annotation, rooted in case relation-based structurization of sentimental knowledge.

## 2. SENTIMENT STRUCTURIZATION

A sentiment analysis system should be capable of understanding and extracting the aspectual sentiments present in a text. The currently most widely used structures for sentiment extraction are *Holder* [3], *Topic* [5], and other domain dependent attributes [4]. However, real life users are not always interested in all sentimental aspects at the same time, but rather look for opinion/sentiment changes of some person (“Who”) during some time period (“When”) and depending upon “What”, “Where” or “Why” something happened. Following this hypothesis, the 5W (Who, What, When, Where and Why) constituent extraction technique for sentiment/opinion structurization was introduced [2]. The 5W structure is domain independent and more generic than the existing semantic constituent extraction structures.

In order to evaluate 5W extraction, three annotators were asked to annotate 5Ws in terms of noun chunks in a corpus of 2,234 Bengali sentences. They were instructed to find the principle opinionated verb in a sentence and successively extract 5W components by asking 5W questions to the verb. The inter-annotator agreements for the identification of subjective sentences for opinion summary between all three annotators are reported in Table 1. The agreement of tag values for each of the 5Ws between two of the annotators (X and Y) are listed in Table 2. As can be seen in the tables, in the present task the inter-annotator agreement is better for Who, When and Where annotation than for What and Why (even though only a small number of documents were considered). Sentiment tagging is always very ambiguous because it changes from the writer’s to the reader’s perspective. Therefore it is very hard to achieve high agreement scores on sentiment data. However, it is important to note that 5W annotation is quite fast: annotation is a vital, albeit tedious task, but 5W annotation is easy to adopt to a new language.

**Table 1: Sentence level inter-annotator agreement**

Annotators:	X-Y	X-Z	Y-Z	All
Agreement:	73.87%	69.06%	60.44%	58.66%

**Table 2: 5W level inter-annotator agreement**

Tag:	Who	What	When	Where	Why
X-Y:	88.45%	64.66%	76.45%	75.23%	56.23%

**Table 3: Sentence level 5W co-occurrence patterns**

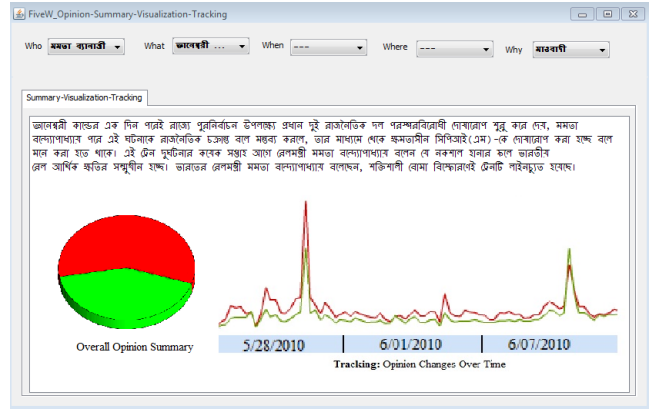
Tag	Who	What	When	Where	Why
Why	28.33%	64.91%	23.66%	12.02%	
Where	78.01%	70.63%	48.63%		
When	73.34%	62.89%			
What	58.56%				
Overall	73.50%	64.23%	57.23%	68.65%	32.00%

Further discussion with the annotators revealed that the psychology of the annotators was to try to locate all 5Ws in every sentence, whereas in general not all 5Ws are present in each sentence. Furthermore, each W type is normally present only once in a sentence, but may sometimes appear twice. As observed, the most ambiguous tag to identify is “Why”. Table 3 presents the sentence level co-occurrence patterns of the 5Ws in the corpus. The 5Ws do not appear together regularly in the corpus. Hence sequence labeling with 5W tags using any machine learning techniques will lead to a label bias problem, and may not be an acceptable solution.

### 3. SENTIMENT VISUALIZATION

In order to investigate how well the 5W annotation method fulfills the needs of the end users, a multi-genre 5W constituent based textual summarization, visualization and tracking system was developed [1]. The system allows users to generate visual sentiment tracking with a polarity wise graph, according to any dimension or combination of the 5W dimensions. A snapshot of the system is presented in Figure 1. The user interface includes five drop down boxes, giving options for presenting each 5W dimension. In essence, the system is a multi-document extractive opinion summarizer for Bengali: documents are preprocessed by a subjectivity classifier which marks the subjective sentences and a polarity classifier which marks the evaluative expressions with their polarity.

All constituents extracted from each sentence are accumulated and the union of all the constituents is treated as the document level sentiment theme. After constituent extraction, a constituent clustering algorithm partitions the set of documents into a finite number of groups or clusters in terms of 5W opinion constituents. Documents are represented as a vector of 5W constituents present in the opinionated sentences within the document. The similarity between vectors is calculated by assigning numerical weights to 5W opinion constituents and then using the Cosine Similarity Measure.

**Figure 1: The 5W sentiment visualization system**

The document clusters are formed as tightly coupled networks where the nodes are the extracted sentiment constituents and the edges represent the relationships between them. The relevant sentences from each constituent cluster are then extracted. The summarization system is dynamic and the output depends on user dimension choices: the most “informed” sentences (those that reflect the concise contextual content of the cluster) are identified, based on the assumption that a text fragment in a document is relevant if it is highly related to many relevant text fragments of other documents in the same cluster. The basic idea is to use the Page Rank algorithm to include all the constituents’ nodes by finding the shortest path which covers all the user’s desired constituents’ nodes while maximizing the accumulated edge scores.

The notion of the 5Ws is well-rooted in Panini’s *karaka* (case relation) theory (~500 BC). We believe that annotation with 5W constituents allows for better text understanding as well as user-intuitive sentiment visualization and tracking, while the method in itself caters for fast and easy annotation.

### 4. REFERENCES

- [1] A. Das, S. Bandyopadhyay, and B. Gambäck. The 5W structure for sentiment summarization-visualization-tracking. *13th Int. Conf. on Intelligent Text Processing and Computational Linguistics*, pp. 540–555, New Delhi, India, 2012. Springer.
- [2] A. Das, A. Ghosh, and S. Bandyopadhyay. Semantic role labeling for Bengali noun using 5Ws: Who, What, When, Where and Why. *6th Int. Conf. on Natural Language Processing and Knowledge Engineering*, pp. 1–8, Beijing, China, 2010. IEEE.
- [3] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. *20th Int. Conf. on Computational Linguistics*, pp. 1367–1373, Geneva, Switzerland, 2004. ACL.
- [4] N. Kobayashi, R. Iida, K. Inui, and Y. Matsumoto. Opinion extraction using a learning-based anaphora resolution technique. *2nd Int. Joint Conf. on Natural Language Processing*, pp. 175–180, Jeju Island, Republic of Korea, 2005. Springer.
- [5] L.-W. Ku, L.-Y. Lee, T.-H. Wu, and H.-H. Chen. Major topic detection and its application to opinion summarization. *28th Int. Conf. on Research and Development in Information Retrieval*, pp. 627–628, Salvador, Brazil, 2005. ACM SIGIR.