Sentiment Analysis: What is the End User's Requirement?

Amitava Das
Department of Computer and Engineering
Jadavpur University
Kolkata-700032, India
amiava.santu@gmail.com

Sivaji Bandyopadhyay
Department of Computer and Engineering
Jadavpur University
Kolkata-700032, India
sivaji_cse_ju@yahoo.com

Björn Gambäck
IDI, Norwegian University of Science and Technology (NTNU)
Trondheim, Norway
gamback@idi.ntnu.no

ABSTRACT
In this paper we address the Sentiment Analysis problem from the end user's perspective. An end user might desire an automated at-a-glance presentation of the main points made in a single review or how opinion changes time to time over multiple documents. To meet the requirement we propose a relatively generic opinion 5Ws structurization, further used for *textual and visual summary and tracking*. The 5W task seeks to extract the semantic constituents in a natural language sentence by distilling it into the answers to the **5W** questions: **Who**, **What**, **When**, **Where** and **Why**. The visualization system facilitates users to generate sentiment tracking with textual summary and sentiment 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***".

Keywords
Sentiment Summarization, Sentiment Visualization, Sentiment Tracking, 5W Structurization

1. Necessity is the Mother of All Invention
In today's digital age, text is the primary medium of representing and communicating information, as evidenced by the pervasiveness of e-mails, instant messages, documents, weblogs, news articles, homepages and printed materials. Our lives are now saturated with textual information, and there is an increasing urgency to develop technology to help us manage and make sense of the resulting information overload. While expert systems have enjoyed some success in assisting information retrieval, data mining, and natural language processing (NLP) systems, there is a growing requirement of Sentiment Analysis (SA) system, can process automatically the plethora of sentimental information available in online electronic text.

WIMS'12, June 13-15, 2012 Craiova, Romania

So far the Sentiment Analysis research becomes quite mature after a few decade of cultivation. The focus of this paper is on aggregating and representing sentiment information drawn from an individual document or from a collection of documents. Sentiment/opinion aggregation is necessary requirement at the end user's perspective. For example, an end user might desire an at-a-glance presentation of the main points made in a single review or how opinion changes time to time over multiple documents. On real-life applications, to provide a completely automated solution is the ultimate desired goal of all the sentiment analysis research. An intelligent system should smart enough to aggregate all the scattered sentimental information from the various blogs, news article and from written reviews.

The role of any automatic system is to minimize human user's effort and produce a good sensible output.

There is no doubt that aggregation of sentiment is necessary but it is very hard to find out the consensus among researchers that how the sentimental information should be aggregated. Although a few systems like Twitter Sentiment Analysis Tool[1], TweetFeel[2] are available in World Wide Web (WWW) since last few years still more research efforts are necessary to match the user satisfaction level and the social need. We try to address this issue with the light of previous works in the section 2. As there is no consensus thus we experimented with multiple output genres like: multi-document 5W variable constituent based textual summarization with overall visualization and tracking. There are various technical challenges to develop the 5W extraction systems: are addressed in the section 5 along with the proposed solution architecture. The details of the developed textual summarization system are discussed in the section 7. The visualization and tracking systems are discussed in the section 8. The paper is finally concluded with section 9.

2. Opinion Summary: Topic-Wise, Polarity-Wise or Other-Wise?
Aggregation of information is the necessity from the end user's perspective but it is nearly impossible to make consensus about the output format or how the data should be aggregated. Researchers tried with various types of output format like textual or visual summary or overall tracking with time dimension. The next key issue is "*how the data should be aggregated?*". Dasgupta and Ng [1] throw an important question: "***Topic-wise, Sentiment-wise, or Otherwise?***" about the opinion summary generation techniques. Actually the output format varies on end user's requirement and the domain, the

[1] http://twittersentiment.appspot.com/

[2] http://www.tweetfeel.com/

system made for. Instead of digging for the answer of the unresolved debate we experimented with multiple output formats. At first we will look into the topic-wise, polarity-wise and other-wise summarization systems proposed by various previous researchers and then will describe the systems developed by us.

2.1 Topic-Wise

There is clearly a tight connection between extraction of topic-based information from a single document and topic-based summarization of that document, since the information that is pulled out can serve as a summary; see [2] for a brief review (Section 5.1). Obviously, this connection between extraction and summarization holds in the case of sentiment-based summarization, as well. We will now glimpse to the various topic-opinion summarization systems, proposed by the previous researchers.

Leveraging existing topic-based technologies is the most common practice for sentiment summarization. One line of practice is to adapt existing topic-based multi-document summarization algorithms to the sentiment setting. Sometimes the adaptation consists simply of modifying the input to these pre-existing algorithms. For instance, [3] propose that one apply standard multi-document summarization to a sub-collection of documents that are on the same topic and that are determined to belong to some relevant genre of text, such as "*argumentative*".

Pang and Lee, [4] propose a two-step procedure for polarity classification for movie reviews, wherein they first detect the objective portions of a document (e.g., plot descriptions) and then apply polarity classification to the remainder of the document after the removal of these presumably uninformative portions. Importantly, instead of making the subjective-objective decision for each sentence individually, they postulate that there might be a certain degree of continuity in subjectivity labels (an author usually does not switch too frequently between being subjective and being objective), and incorporate this intuition by assigning preferences for pairs of nearby sentences to receive similar labels. All the sentences in the document are then labeled as being either subjective or objective through a collective classification process, where this process employs a reformulation of the task as one of finding a *minimum s-t cut* in the appropriate graph.

Ku et al., [5] present techniques for automatic opinion summarization based on topic detection. The system selects representative words from a document set to identify the main concepts in the document set. A term is considered to represent a topic if it appears frequently across documents or in each document. The authors use many weighting mechanism to detect the representative words (topic words) at sentence, paragraph or document level. The identified topic then further used for opinion summarization.

Zhou et al. [6] have proposed the architecture for generative summary from blogosphere. Typical multi-document summarization (MDS) systems focus on content selection followed by synthesis by removing redundancy across multiple input documents. Due to the complex structure of the dialogue, similar subtopic structure identification in the participant-written dialogues is essential. Maximum Entropy Model (MEMM) and Support Vector Machine (SVM) have been used with a number of relevant features.

Kawai et al. [7] developed a news portal site called Fair News Reader (FNR) that recommends news articles with different sentiments for a user in each of the topics in which the user is interested. FNR can detect various sentiments of news articles and determine the sentimental preferences of a user based on the sentiments of previously read articles by the user.

2.2 Polarity-Wise

Indeed the topic-opinion model is the most popular one but there could be a requirement at the end user's perspective that they might look into an at-a-glance presentation of opinion-oriented summary. For example: One market surveyor from company *A* might be interested in the root cause for why their product *X* (suppose camera) become less popular day by day. And for this particular case *A* may wants look into for the *negative* reviews only. Therefore opinion-oriented summary is the end user's requirement here. Relatively a few research efforts could be found on the polarity-wise summarization in the literature than the popular topic-opinion model. We present four important related works here which are significant in both the aspects: problem definition and solution architecture with best of our knowledge.

Hu, [8] developed a review mining and summarization system, works in three steps: (1) mining product features that have been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. Instead of the most popular topic-sentiment model the authors opt for a feature based polarity wise summarization system.

A multi-knowledge based approach for review mining and summarization is proposed by [9] which integrates WordNet, statistical analysis and movie knowledge. After identifying all valid above mentioned feature the final summary has been generated according to the following steps. All the sentences that express opinions on a feature class are collected. Then, the semantic orientation of the relevant opinion in each sentence is identified. Finally, the organized sentence list is shown as the summary. The following is an example of the polarity class wise summary produced by the system.

Das and Chen, [10] develop a methodology for extracting small investor sentiment from stock message boards. Altogether five machine learning classifiers have been used for opinion polarity classification and the final output comprises different classifier algorithms coupled together by a voting scheme. The five classifier algorithms rely on a different approach to message interpretation. Some of them are language independent, and some are not. The final classification is based on achieving a simple majority vote amongst the five classifiers, i.e., three of five classifiers should agree on the message type. Finally the summary generated statistically for each polarity classes such as *bullish*, *bearish* or *neutral*.

2.3 Visualization

To convey all the automatically extracted knowledge to the end user concisely the graphical or visualized output format is one of the trusted and well acceptable methods. Thus a numbers of researcher tried to leverage the existing or newly developed graphical visualization methods for the opinion summary presentation. We describe some noteworthy related previous works on opinion summary visualization techniques.

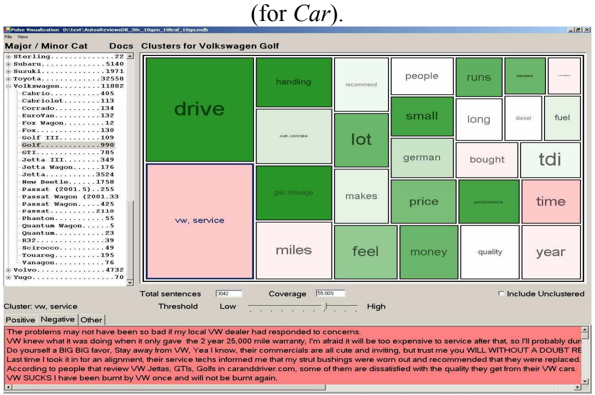
Gamon et al., [11] present a system called Pulse, extracts topic-sentiment from customer written review, which is in general free text. Pulse also displays the extracted information into two dimensions, i.e. topic and sentiment, simultaneously.

As example the Pulse visualization has been reported in the Figure 1 for the topic “*car*”.

Yi and Niblack, [12] proposed several methodologies for sentiment extraction and visualization using *WebFountain* [13]. A *Sentiment Miner* system has been developed with the basic backbone architecture of the WebFountain. The *Sentiment Miner* system is trained for both the structured and unstructured application-specific data. A topic-sentiment model has been followed with the following hypothesis.

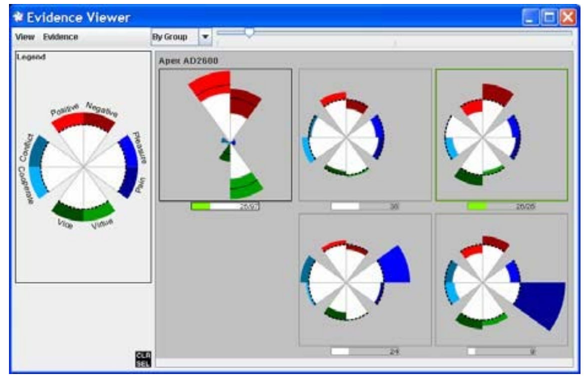
Carenini et al. [14][3] present and compare two approaches to the task of multi document opinion summarization on evaluative texts. The first is a sentence extraction based approach while the second one is a natural language generation-based approach. Relevant extracted features are categorized in two types: User Defined Features (UDF) and Crude Features (CF) as described in [8].

Figure 1: Screenshot of the Pulse user interface showing the taxonomy and the Tree Map with labeled clusters and sentiment coloring, and individual sentences from one cluster (for *Car*).



The research effort by [15] present techniques to extract and visualize the affective content of documents and describes an interactive capability for exploring emotion in a large document collection. The proposed system first automatically identifies affective text by comparing each document against a lexicon of affect-bearing words and obtains an affect score for each document. A number of visual *metaphors* have been proposed to represent the affect in the collection and a number of tools that can be used to interactively explore the affective content of the data. The visualization has been enhanced to support visual analysis of sentiment as shown in the Figure 2.

Figure 2: Affect summary and variation for “service” cluster and “picture” cluster.



2.4 Tracking

In many applications, analysts and other users are interested in tracking changes in sentiment about a product, political candidate, company or other issues over time. The tracking system could be a good measure to understand the people’s sentiment changes or it could be helpful sociological survey also. In general sense tracking means plotting of sentiment values over time into a graphical visualization. We mention some significant research efforts on opinion tracking.

The Lydia[4] project (also called TextMap) [16] seeks to build a relational model of people, places and many more other things through natural language processing of news sources and the statistical analysis of entity frequencies and co-locations. The system track the temporal and spatial distribution of the entities in the news: *who* is being talked about, by *whom*, *when* and *where*? The Lydia system relies on visual output and the previously mentioned aspect are reported by the *juxtapositional*, *spatial* and the *temporal* entity analysis.

Ku et al., [17] hypothesize that opinion extraction, opinion summarization and opinion tracking are three important techniques for understanding opinions. Opinion extraction mines opinions at word, sentence and document levels from articles. Opinion summarization summarizes opinions of articles by telling sentiment polarities, degree and the correlated events. Opinion tracking visually reports the opinion changes over time. The authors investigated their techniques on both the news and web blog articles. TREC[5] and NTCIR[6] articles are collected from the web blogs, serve as the information sources for this task.

Mishne and Rijke, [18] demonstrate a system for tracking and analyzing moods of bloggers worldwide. The demonstrated system is trained on the largest blogging community, LiveJournal[7]. Users of LiveJournal, currently the largest weblog community, have the option of reporting their mood at the time of the post; users can either select a mood from a predefined list of 132 common moods such as “amused” or “angry,” or enter free-text. The authors developed a system, called MoodViews[8], a collection of tools for analyzing, tracking and visualizing moods and mood changes in blogs posted by LiveJournal users. An example is shown in Figure 3.

---

[3] http://www.cs.ubc.ca/~carenini/storage/SEA/demo.html

[4] http://www.textmap.com/

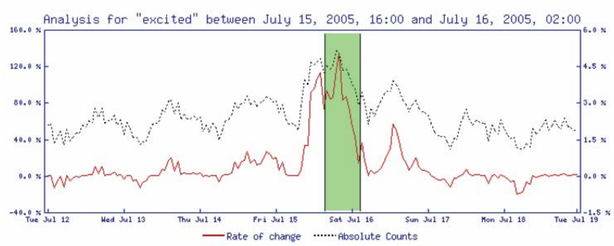
[5] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

[6] http://research.nii.ac.jp/ntcir/index-en.html

[7] http://www.livejournal.com/

[8] http://moodviews.com/

Figure 3: Moodsignals uncovering the excitement peak on July 16, 2005: the release of a new Harry Potter book.



Fukuhara et al., [19] present their research effort on finding the temporal sentiment analysis that analyzes temporal trends of sentiments and topics from a text archive. The system accepts texts with timestamp such as Weblog and news articles, and produces two kinds of graphs, i.e., (1) ***topic graph*** that shows ***temporal change*** of topics associated with a sentiment, and (2) ***sentiment graph*** that shows ***temporal change*** of sentiments associated with a topic.

3. The Accountability of the Present 5W Sentiment System

We mentioned a few (Due to space complexity) of noteworthy related works in this section. During the literature survey we realized that there is no consensus among the researchers could be found on the output format of any sentiment summarization system.

Instead of digging for the answer of the unresolved debate we experimented with multiple output formats: multi-document topic-opinion textual summary but realizing the end user's requirement and to less their effort and to present an *at-a-glance* representation we devise a 5W constituent based textual summarization, visualization and tracking system. The 5W constituent based summarization system is a multi-genre system. The system facilitates users to generate sentiment tracking with textual summary and sentiment 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*". During the related work discussion we categorize the previous systems in "*Topic-Wise*", "*Polarity-Wise*" or "*Other-Wise*" genres. In the "*Other-Wise*" genre we described the necessity of the visualization and tracking systems. As par our understanding the 5W constituent based summarization system fall into every genre and the supportive argumentations from our side are as follows:

Topic-Wise: The 5W system facilitates users to generate sentiment summary based on any customized topic like Who, What, When, Where and Why and based on any dimension or combination of dimensions as they want.

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

Visualization and Tracking: The visualization 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 been generated with a timeline.

From the next section we describe the development process of our 5W constituent based textual and visual summarization and tracking system.

4. Resource Organization

The present system has been developed for the Bengali language. Resource acquisition is one of the most challenging obstacles to work with resource constrained languages like Bengali. Bengali is the fifth popular language[9] in the World, second in India and the national language in Bangladesh. Extensive NLP research activities in Bengali have started recently but resources like annotated corpus, various linguistic tools are still unavailable for Bengali. The manual annotation of the gold standard Bengali corpus is described in following section. The most effective features are chosen experimentally. All the features that have been used to develop the present system are described in Feature Organization section.

4.1 Corpus

The details of corpus development could be found in [20] for Bengali. We obtained the corpus from the authors. For the present task a portion of the corpus from the editorial pages, i.e., Reader's opinion section or Letters to the Editor Section containing 28K word-forms have been manually annotated with sentence level opinion constituents. The detail statistics about the corpus is reported in Table 1.

Table 1: Bengali News Corpus Statistics

Statistics	NEWS
Total number of documents	100
Total number of sentences	2234
Avgerage number of sentences in a document	22
Total number of wordforms	28807
Avgerage number of wordforms in a document	288
Total number of distinct wordforms	17176

4.2 Annotation

Annotators were asked to annotate 5Ws in Bengali sentences in terms of Bengali noun chunks. Instructions have been given to annotators to find out the principle opinionated verb in a sentence and successively extract 5W components by asking 5W questions to the principle verb.

Table 2: Agreement of annotators at each 5W level

Tag	Annotators X and Y Agree percentage
Who	88.45%
What	64.66%
When	76.45%
Where	75.23%
Why	56.23%

Table 3: Agreement of annotators at sentence level

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Avg.

9 http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

Percentage	73.87%	69.06%	60.44%	67.8%
All Agree	58.66%			

The agreement of annotations between two annotators (Mr. X and Mr. Y) has been evaluated. The agreements of tag values at each 5W level are listed in Tables 2. For the evaluation of the extractive summarization system gold standard data has been prepared and three annotators took part. The inter-annotator agreement for the identification of subjective sentences for opinion summary is reported in Table 3.

It has been observed that in the present task the inter-annotator agreement is better for Who, When and Where level annotation rather than What and Why level though a small number of documents have been considered.

Further discussion with annotators reveals that the psychology of annotators is to grasp all 5Ws in every sentence, whereas in general all 5Ws are not present in every sentence. But the same groups of annotators are more cautious during sentence identification for summary as they are very conscious to find out the most concise set of sentences that best describe the opinionated snapshot of any document. The annotators were working independent of each other and they were not trained linguists. As observed, the most ambiguous tag to identify is "Why". To understand the distribution pattern of 5Ws in a corpus we gather a statistics for each 5W tag level as listed in Table 4.

Table 4: Sentence wise co-occurrence pattern of 5Ws

Tags	Percentage					
	Who	What	When	Where	Why	Overall
Who	-	58.56%	73.34%	78.01%	28.33%	73.50%
What	58.56%	-	62.89%	70.63%	64.91%	64.23%
When	73.34%	62.89%	-	48.63%	23.66%	57.23%
Where	78.0%	70.63%	48.63%	-	12.02%	68.65%
Why	28.33%	64.91%	23.66%	12.02%	-	32.00%

Another important observation is that 5W annotation task takes very little time for annotation. Annotation is a vital tedious task for any new experiment, but 5W annotation task is easy to adopt for any new language.

4.3 Subjectivity Classifier

The subjectivity classifier as described in [21] has been used. The recall measure of the present classifier is greater than its precision value. The evaluation results of the classifier are 72.16% (Precision) on the NEWS Corpus.

4.4 Polarity Classifier

The polarity identifier chosen for the present task is described in [22]. It has a good accuracy level.

4.5 Dependency Parsers

Shallow parsers[10] for Indian languages developed under a Government of India funded consortium project named Indian Language to Indian Language Machine Translation System (IL-ILMT) are now publicly available.

[10] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

5. The 5W Extraction

The 5Ws semantic role labeling task demands and addressing various NLP issues such as: predicate identification, argument extraction, attachment disambiguation, location and time expression recognition. To solve these issues the present system architecture relies on Machine Learning technique and rule-based methodologies simultaneously.

One of the most important milestones in SRL literature is CoNLL-2005 Shared Task[11] on Semantic Role Labeling. All most all SRL research group participated in the shared task. System reports of those participated systems eminently prove that Maximum Entropy[12] (ME) based models work well in this problem domain as 8 among 19 systems used ME as the solution architecture. The second best performing system [23] uses ME model uses only syntactic information without using any pre or post processing.

Table 4 presents the distribution pattern of 5Ws in overall corpus. It is very clear that 5Ws are not very regular jointly in the corpus. Hence sequence labeling with 5Ws tags using ME will lead a label biased problem (as we reported in Section 7) and may not be an acceptable solution for present problem definition as concluded in [23] (although in a different SRL task).

We apply both rule-based and statistical techniques jointly to the final system. The rules are being captured by acquired statistics on training set and linguistic analysis of standard Bengali grammar. The features used in the present system are reported in the following section.

5.1 The Feature Organization

The features to be found most effective are chosen experimentally. Bengali is an electronically resource scarce language, thus our aim was to find out the less number of features but the features should be effective. Involving more number of features will demand more linguistic tools, which are not readily available for the language. All the features that have been used to develop the present system are categorized as Lexical, Morphological and Syntactic features. These are listed in the Table 5 below and have been described in the subsequent subsections. The tool has been used here is Bengali Shallow Parser[13] developed under Indian Languages to Indian Languages machine Translation (IL-ILMT) project.

Table 5: Features

Types	Features	
Lexical	POS	
	Root Word	
Morphological	**Noun**	Gender
		Number
		Person
		Case
	Verb	Voice
		Modality
Syntactic	Head Noun	
	Chunk Type	

[11] http://www.lsi.upc.es/~srlconll/st05/st05.html

[12] http://maxent.sourceforge.net/

[13] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

	Dependency Relations

5.1.1 Part of Speech (POS)

POS of any word cannot be treated as direct clue of its semantic but it definitely helps to identify it. Finding out the POS of any word can reduce the search space for semantic meaning. It has been shown by [24], [25] etc. that the part of speech of any word in sentences is a vital clue to identify semantic role of that word.

5.1.2 Root Word

Root word is a good feature to identify word level semantic role especially for those types of 5Ws where dictionaries have been made like “When”, “Where” and “Why”. There are various conjuncts and postpositions, which directly indicate the type of predicate present in any sentence. As example জন্য, হেতু give clue that the next predicate is causative (“Why”).

5.1.3 Nouns

5.1.3.1 GENDER

Gender information is essential to relate any chunk to the principle verb modality. In the case of “What”/”Whom” ambiguities gender information help significantly. For inanimate objects it will be null and for animates it has definitely a value. Bengali is not a gender sensitive language hence this feature is not such significant linguistically rather number and person features. But the statistical co-occurrence of gender information with the number and person information is significant.

5.1.3.2 NUMBER

Number information help to identify specially for “Who”/”What” ambiguities. As we reported in inter-annotator agreement section “Who” has been identified first by matching modality information of principle verb with corresponding number information of noun chunks.

5.1.3.3 PERSON

Person information is as important as number information. It helps to relate any head of noun chunks to principle verb in any sentence.

5.1.3.4 CASE

Case markers are generally described as *karaka* relations of any noun chunks with main verb. It has been described that semantically *karaka* is the ancestor of all semantic role interpretations. Case markers are categorized as Nominative, Accusative, Genitive and Locative. Case markers are very helpful for almost in every 5W semantic role identification task.

5.1.4 Verb

5.1.4.1 Voice

The distinction between active and passive verbs plays an important role in the connection between semantic role and grammatical function, since direct objects of active verbs often correspond in semantic role to subjects of passive verbs as suggested by various researchers [24]. A set of hand-written rules helps to identify the voice of any verb chunk. The rules rely on presence auxiliary verbs like হয়েছে, হোক etc. indicate that the main verb in that particular chunk is in passive form. In Bengali active voice sentences are generally dropped copula like: In the example sentence the verb হয় is dropped.

```
This is my umbrella.
এটা (null) আমার ছাতা ।
```

5.1.4.2 Modality

Honorific markers are very distinctly used in Bengali and it directly reflects by the modality marker of any verb. As example the honorific variation করা/do are as কর (used with তুই: 2nd person either of same age or younger), করো (used with তুমি: 2nd person either of same age or slightly elder) and করুন (used with আপনি: 2nd persond generally for aged or honorable person). Verb Modality information helps to identify especially the “Who” tag. “Who” is identified first by matching modality information of principle verb with corresponding number information of noun chunks.

5.1.5 Head Noun

The present SRL system identifies chunk level semantic roles. Therefore morphological features of chunk head is only important rather other chunk members. Head words of noun phrases can be used to express selectional restrictions on the semantic role types of the noun chunks. For example, in a communication frame, noun phrases headed by Ram, brother, or he are more likely to be the SPEAKER (Who), while those headed by proposal, story, or question are more likely to be the TOPIC (What).

5.1.6 Chunk Type

Present SRL system identifies noun chunk level semantic roles. Hence chunk level information is effectively used as a feature in supervised classifier and in rule-based post processor.

5.1.7 Dependency Relations

It has been profoundly established that dependency phrase-structures are most crucial to understand semantic contribution of every syntactic nods in a sentence [24], [25]. A statistical dependency parser has been used for Bengali as described in [26].

5.2 Using MEMM

MEMM treats 5Ws semantic role labeling task as a sequence tagging task. MEMM acquires symbolic patterns that rely on the syntax and lexical semantics and morphological features of a phrase head. With the all selected features properties in mind and supported by series of experimentation, we finalize the final features (described in Table 5) for each chunk level in an input sentence. For pedagogical reasons, we may describe some of the features as being multi-valued (e.g. root word) or categorical (e.g. POS category) features. In practice, however, all features are binary for the MEMM model. In order to identify features we started with Part Of Speech (POS) categories and continued the exploration with the other features like chunk, Dependency relation and morphological features. The feature extraction pattern for any Machine Learning task is crucial since proper identification of the entire features directly affect the performance of the system. 5Ws Semantic role labeling is difficult in many ways. A sentence does not always contain all 5Ws. Although Bengali is defined as a verb final language but there is no certain order in occurrence among these 5Ws in a sentence. The performance of 5W SRL task by MEMM is reported in Table 9.

It is noticeable that the performance of the MEMM-based model differs tag-wise. While precision values for “Who”, “When” and “Where” is good but recall yielded i.e. system failed to identify in various cases. In “What” cases system identified most of the cases as recall is high but also make so many false hits as precision is low. For tag label “Why” precision and recall values both are low as reported.

For such heterogeneous problem nature we propose a hybrid system as rule-based post processor followed by Machine Learning. The rule-based post processor can identify those cases missed by ML method and can reduce false hits generated by statistical system. These rules are formed by heuristic on gold standard data and standard Bengali grammar.

5.3 Rule-Based Post-Processing

As described earlier post-processing is necessary in this setup. The rules developed here are either based on syntactic grammar, manually augmented dictionary or corpus heuristic. Rules for each tag label are described in the next sub-sections.

5.3.1 Who? Who was involved?

As described earlier system failed to identify “Who” in many cases. As an example:

নিমন্ত্রিত না হলেও তোমার/Who যাওয়া উচিত ছিল সেখানে।

Though you are not invited but you/Who should go there.

System fails in this type of cases, because the targeted chunk head is a pronoun and it is situated at almost in the middle of the sentence whereas “Who” is generally situated at initial positions in a sentence as Bengali is a Verb final and Subject initial language. Moreover the system made some false hits too. As an example:

দরজাটা বন্ধ করে দেওয়াই ভালো।

Closing the door may be the best option.

In the previous case system mark দরজাটা/door as a “Who” whereas the “Who” is “you” (2nd person singular number), which is absent. This is a perfect example of the label-bias problem. The system is biased towards those chunks at initial position of sentences. We developed rules using case marker, Gender-Number-Person (GNP), morphological subject and modality features to disambiguate these types of phenomena. These rules help to stop false hits by identifying that no 2nd person phrase was there in the type of second example sentences and empower to identify proper phrases by locating proper verb modality matching with the right chunk.

5.3.2 What? What happened?

We make use of only positional information for “What” or object identification. There is less syntactic, orthographic and morphological difference between “Who” and “What”. For that reason a reduction methodology has been used as “Who” has been detected by system first and “What” has been tagged among rest of the noun chunks with positional factor in the sentence.

5.3.3 When? When did it take place?

Table 6: Categories of Time Expressions

	Bengali	English Gloss
General	সকাল/সন্ধ্যে/রাত/ভোর...	Morning/evening/night/daw
	টার সময়/ঘটিকায়/মিনিট	O clock/hour/minute
	সোমবার/মঙ্গলবার/রবিবার	Monday/Tuesday/Sunday
	বৈশাখ/জৈষ্ঠ/...	Bengali months…
	জানুয়ারী/ফেব্রুয়ারী	January/February…
	দিন/মাস/বছর...	Day/month/year…
	কাল/ক্ষন/পল...	Long time/moment…
Relative	আগে/পরে...	Before/After…
	সামনে/পেছনে...	Upcoming/
Special Cases	উঠলে/থামলে	When rise/When stop…

Time expressions could be categorized in two types as General and Relative as listed in Table 6. In order to apply rule-based post-processor we developed a manually augmented list with pre-defined categories as described in Table 6.

5.3.4 Where? Where did it take place?

Identification of “Where” simply refers to the task of identification locative marker in NLP. Similar to “When”, we categorized “Where” as general and relative as listed in Table 7.

Rules have been written using a manually edited list as described in Table 7. Morphological locative case marker feature have been successfully used in the identification of locative marker.

Table 7: Categories of Locative Expressions

	Bengali	English Gloss
General	মাঠে/ঘাটে/রাস্তায়	Field/Ghat/Road
Relative	আগে/পরে...	Before/After…
	সামনে/পেছনে...	Front/Behind

5.3.5 Why? Why did it happen?

The particular constituent role assignment for “Why” is the most challenging task as it is separately known as the opinion argument identification. As reported in previous sections inter-annotator agreement and overall distribution regularity is very low. For irregular and small occurrence of “Why” leads poor result in ML-based technique. Inter-annotator agreement shows that even human annotators also disagree about the “Why” tag. To resolve this problem we need a relatively large corpus to learn fruitful feature similarities among argument structures.

A manually generated list of causative postpositional words and pair wise conjuncts as reported in Table 8 has been prepared to identify argument phrases in sentences.

Table 8: Categories of Causative Expressions

	Bengali	English Gloss
General	জন্য/কারনে/হেতু...	Hence/Reason/Reason
Relative	যদি_তবে	If_else
	যদিও_তবুও	If_else

6. Performance of 5Ws Extraction

The performance result of ML (1) technique has been reported in Table 9. After using rule-based postprocessor the system (2) performance increases as listed in the following Table 9.

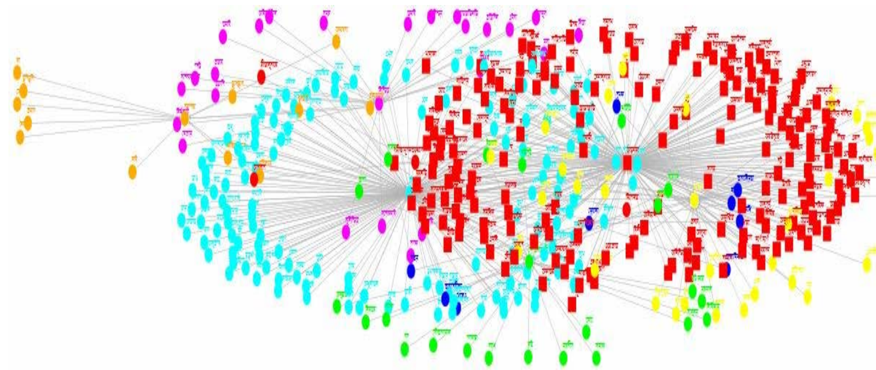
Table 9: Performance of 5Ws Opinion Constituents by MEMM + Rule Based-Post Processing

Tag	Precision		Recall (%)		F-measure		Avg. F-	
	1	2	1	2	1	2	1	2
Who	76.2	79.6	64.3	72.6	69.8	75.9	62.2	68.1
What	61.2	65.5	51.3	59.6	55.9	62.4		
When	69.2	73.4	58.6	66.0	63.4	69.5		
Wher	70.0	77.7	60.0	69.7	64.6	73.4		
Why	76.2	63.5	53.9	55.6	57.4	59.2		

7. The Summarization System

The present system is a multi-document extractive opinion summarization system for Bengali. Documents are preprocessed with the subjectivity identifier (described in section 4.3) followed by the polarity classifier (described in section 4.4).

Figure 4: Document Level Theme Relational Graph by NodeXL



The subjectivity classifier the subjective sentences from documents and the polarity classifier marks the evaluative expression with its polarity. All the constituents extracted from each sentence are accumulated and the unions of all the constituents are treated as the document level sentiment theme.

After the constituent extraction documents are clustered depending upon common constituents at document level. The document clusters are then formed as tightly coupled network. The node of the network is the extracted sentiment constituents and the edges represent the relationship among them.

The next major step is to extract relevant sentences from each constituent cluster that reflects the contextual concise content of the current constituent cluster. Our summarization system is a dynamic one and the output depends on user's dimension choices. To adopt this kind of special need we used Information Retrieval (IR) based technique to identify the most "**informed**" sentences from the constituents cluster and it can be termed as IR based cluster center for that particular cluster. With the adaptation of ideas from page rank algorithms [27], it can be easily observed that a text fragment (sentence) 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 cover all the constituents' node in the network by the shortest path algorithm as given by user. The page rank algorithm helps to find out the shortest distance which covers all the desired constituents' node and maximizes the accumulated edge scores among them. Accordingly sentences are chosen based on the presence of those particular constituents. The detail description could be found in the following subsection.

7.1 Constituent based Document Clustering

Constituent clustering algorithms partition a set of documents into 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 into various subjective sentences.

The similarity between vectors is calculated by assigning numerical weights to 5W opinion constituents and then using the cosine similarity measure as specified in the following equation.

$$s\left(\vec{d}_k,\vec{d}_j\right)=\vec{d}_k.\vec{d}_j=\sum_{i=1}^{N} w_{i,k}\times w_{i,j}$$

where $\vec{d}_k$ and $\vec{d}_j$ are the document vectors. N is the total number of unique 5Ws that exist in the document set $\vec{d}_k$ and $\vec{d}_j$. The $W_{i,k}$ and $W_{i,k}$ are the 5W opinion constituents that exist in the documents $\vec{d}_k$ and $\vec{d}_j$ respectively. An example of inter-document theme cluster has been reported in Table 10. The numeric scores are the similarity association value assigned by the clustering technique. A threshold value of greater than 0.5 has been chosen experimentally to construct the inter-document theme relational graph in the next level.

To better aid our understanding of the automatically determined category relationships we visualized this network using the Fruchterman-Reingold force directed graph layout algorithm [28] and the NodeXL network analysis tool [29][14]. A constituent relational model graph drawn by NodeXL is shown in the Figure 4.

Table 10: Theme Clusters by 5W Dimensions

Generated Clusters						
5Ws	**Constituents**	**Doc1**	**Doc2**	**Doc3**	**Doc4**	**Doc5**
Who	Mamata Banerjee	**0.63**	0.01	**0.55**	**0.93**	0.02
	West Bengal CM	0.00	0.12	**0.37**	0.10	0.17
What	Gyaneswari Express	**0.98**	**0.79**	**0.58**	0.47	0.36
	Derailment	**0.98**	**0.76**	0.35	0.23	0.15
When	24/05/2010	**0.94**	0.01	0.01	0.01	0.01
	Midnight	**0.68**	**0.78**	0.01	0.01	0.01
Where	Jhargram	**0.76**	0.25	0.01	0.13	**0.76**
	Khemasoli	**0.87**	0.01	0.01	0.01	0.01
Why	Maoist	**0.78**	**0.89**	0.06	0.10	0.14
	Bomb Blast	0.13	**0.78**	0.01	0.01	**0.78**

7.2 Sentence Selection for the Summary

The present system is an extractive opinion summarization. The major step is to extract relevant sentences from each constituent cluster that reflects the contextual concise content of relevant cluster. First the page rank algorithm [27] finds out the shortest distance which covers all the desired constituents' node and maximizes the accumulated edge scores among them and

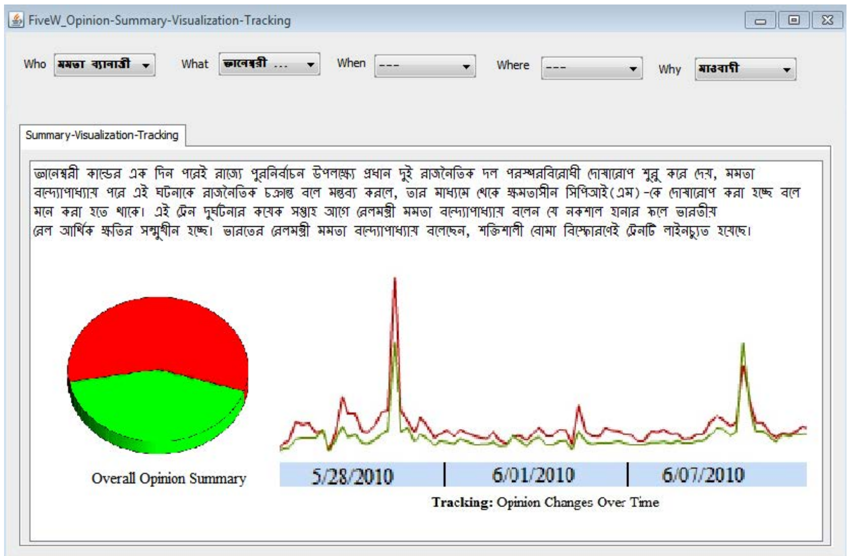
[14] Available from http://www.codeplex.com/NodeXL

accordingly sentences are chosen which cover all the desired nodes.

8. The Visualization and Tracking

The visualization and the tracking system consist of five drop down boxes. The drop down boxes give options for individual 5W dimension of each unique Ws that exist in the corpus. The present visual tracking system facilitates users to generate opinion polarity wise graph based visualization and summary on any 5W dimension and combination of 5W dimensions as they want.

Figure 5: A Snapshot of the Present Summarization System



To generate dimension specific and overall polarity wise graph the system relies on the polarity scores assigned by the polarity identifier. Top ranked 30% of total extracted sentences are chosen as a dimension wise summary. But the graph is built using all scores from all the extracted sentences and ordered time wise by means of "When" tag. A snapshot of the present system has been shown Figure 5.

8.1 Experimental Result

For evaluation we check system identified sentences with every human annotator's gold standard sentences and finally we calculated the overall accuracy of the system as reported in Table 11.

Table 11: Final Results subjective sentence identification for summary

Metrics	X	Y	Z	Avg
Precision	77.65%	67.22%	71.57%	72.15%
Recall	68.76%	64.53%	68.68%	67.32%
F-Score	72.94%	65.85%	70.10%	69.65%

It was a challenge to evaluate the accuracy of the dimension specific summaries. According to the classical theory we should make human extracted set for every dimension combinations, but it is too difficult to develop such gold standard dataset. What we propose here is a direct human evaluation technique.

Two evaluators have been involved in the present task and asked to give score to each system generated summaries. We use a 1-5 scoring technique whereas 1 denotes very poor, 2 denotes poor, 3 denotes acceptable, 4 denotes good and 5 denotes excellent. The final evaluation result of the dimension specific summarization system is reported in Table 12.

Table 12: Human Evaluation on 5W Dimension Specific Summaries

Tags	Average Scores					
	Who	What	When	Where	Why	Overall
Who	-	3.20	3.30	3.30	2.50	3.08
What	3.20	-	3.33	3.80	2.6	3.23
When	3.30	3.33	-	2.0	2.5	3.00
Where	3.30	3.80	2.0	-	2.0	2.77
Why	2.50	2.6	2.5	2.0	-	2.40

9. Conclusion

The present paper started with a very basic question "*What is the End User's Requirement?*". To answer this question we do believe that our proposed 5W Summarization –Visualization-Tracking system could be treated as a qualitative and acceptable solution. To compare our suggestion we presented a vivid description of the previous works. Another self contributory remark should be mentioned that according to best of our knowledge this is the first attempt on opinion summarization or visual tracking for the language Bengali. Moreover the 5W structurization is new to the community and proposed by us.

10. REFERENCES

[1] Sajib Dasgupta and Vincent Ng. Topic-wise, Sentiment wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification. In the conference of the EMNLP, Singapore, 2009.

[2] Bo Pang; Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, Now Publishers, 2008

[3] Yohei Seki, Koji Eguchi, and Noriko Kando. Analysis of multi-document viewpoint summarization using multi-dimensional genres. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, pages 142–145, 2004.

[4] Bo Pang; Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of ACL, 2004

[5] Ku Lun-Wei, Li Li-Ying, Wu Tung-Ho, and Chen Hsin-Hsi. Major topic detection and its application to opinion summarization. In Proceedings of the SIGIR, pages 627–628, 2005.

[6] Zhou Liang and Hovy Eduard. On the summarization of dynamically introduced information: Online discussions and blogs. In AAAI-CAAW. Pages 237–242, 2006.

[7] Kawai Yukiko, Kumamoto Tadahiko, and Katsumi Tanaka. Fair News Reader: Recommending news articles with different sentiments based on user preference. In Proceedings of Knowledge-Based Intelligent Information and Engineering Systems (KES), number 4692 in Lecture Notes in Computer Science, pages 612–622, 2007.

[8] Hu M. and Liu B. 2004. Mining and summarizing-customer reviews. In Proc. of the 10th ACM-SIGKDD Conf., pages 168–177, New York, NY, USA. ACM Press.

[9] Li Zhuang, Feng Jing, Xiaoyan Zhu, and Lei Zhang. Movie review mining and summarization. In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM), 2006.

[10] Sanjiv R. Das and Mike Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the Web. Management Science, 53(9):1375–1388, 2007.

[11] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining customer opinions from free text. In Proceedings of the International Symposium on Intelligent Data Analysis (IDA), number 3646 in Lecture Notes in Computer Science, pages 121–132, 2005.

[12] Jeonghee Yi and Wayne Niblack. Sentiment mining in WebFountain. In Proceedings of the International Conference on Data Engineering (ICDE), 2005.

[13] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a Webfountain: architecture for very large-scale text analytics. IBM Systems Journal, 43(1):64–77, 2004.

[14] Giuseppe Carenini, Raymond T. Ng, and Adam Pauls. Interactive multimedia summaries of evaluative text. In Proceedings of Intelligent User Interfaces (IUI), pages 124–131. ACM Press, 2006.

[15] Michelle L. Gregory, Nancy Chinchor, Paul Whitney, Richard Carter, Elizabeth Hetzler, and Alan Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 23–30, 2006. ACL.

[16] Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. Lydia: A system for large-scale news analysis. In Proceedings of String Processing and Information Retrieval (SPIRE), number 3772 in LNCS, pages 161–166, 2005.

[17] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In AAAI-CAAW, pages 100–107, 2006.

[18] Gilad Mishne and Maarten de Rijke. Moodviews: Tools for blog mood analysis. In AAAI-CAAW, pages 153–154, 2006.

[19] Tomohiro Fukuhara, Hiroshi Nakagawa, and Toyoaki Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In Proceedings of the ICWSM, 2007.

[20] A. Ekbal and S. Bandyopadhyay (2008). A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation (LRE) Journal, Springer, Vol. 42(2), PP.173-182.

[21] A. Das and S. Bandyopadhyay. Subjectivity Detection using Genetic Algorithm, In the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA10), Lisbon, Portugal, August 16-20, 2010.

[22] A. Das and S. Bandyopadhyay (2010). Phrase-level Polarity Identification for Bengali., In International Journal of Computational Linguistics and Applications (IJCLA), Vol. 1, No. 1-2, Jan-Dec 2010, ISSN 0976-0962, Pages 169-182

[23] Aria Haghighi, Kristina Toutanova and Christopher D. Manning. A Joint Model for Semantic Role Labeling. In CoNLL-2005 Shared Task.

[24] Daniel Gildea and Daniel Jurafsky, Automatic Labeling of Semantic Roles, In Association for Computational Linguistics, 2002.

[25] Martha Palmer, Dan Gildea, Paul Kingsbury, The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics Journal, 31:1, 2005.

[26] A. Ghosh, A. Das, P. Bhaskar, S. Bandyopadhyay. Dependency Parser for Bengali: the JU System at ICON 2009. In NLP Tool Contest ICON 2009, December 14th-17th, 2009, Hyderabad.

[27] Larry Page. PageRank: Bringing Order to the Web. Stanford Digital Library Project, talk. August 18, 1997

[28] Fruchterman Thomas M. J. and Reingold Edward M.(1991). Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164.

[29] Smith Marc, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. 2009. Analyzing (social media) networks with NodeXL. In C&T '09: Proc. Fourth International Conference on Communities and Technologies, LNCS. Springer.