Technical Challenges and Design Issues in Bangla Language Processing

M. A. Karim
Old Dominion University, USA

M. Kaykobad
Bangladesh University of Engineering and Technology, Bangladesh

M. Murshed
Monash University, Australia

Information Science REFERENCE
An Imprint of IGI Global

Managing Director: Lindsay Johnston
Editorial Director: Joel Gamon
Production Manager: Jennifer Yoder
Publishing Systems Analyst: Adrienne Freeland
Development Editor: Monica Speca
Assistant Acquisitions Editor: Kayla Wolfe
Typesetter: Alyson Zerbe
Cover Design: Jason Mull

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Technical challenges and design issues in Bangla language processing / M.A. Karim, M. Kaykobad, and M. Murshed, editors.
pages cm
Includes bibliographical references and index.
Summary: This book addresses the difficulties as well as the overwhelming benefits associated with creating programs and devices that are accessible to the speakers of the Bangla language -- Provided by publisher.
ISBN 978-1-4666-3970-6 (hardcover) -- ISBN 978-1-4666-3971-3 (ebook) -- ISBN 978-1-4666-3972-0 (print & perpetual access) 1. Bengali language--Data processing. 2. Computational linguistics. I. Karim, M. A., 1953- editor of compilation. II. Kaykobad, M., 1954- editor of compilation. III. Murshed, M., 1970- editor of compilation.
PK1658.5.T43 2013
491.4'40285--dc23
2012051564

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 15

যন্ত্র-না (Jantra-Na: Not-Machine) Can Only Feel যন্ত্রনা (Jantrana: Pain)!

Amitava Das
Norwegian University of Science and Technology, Norway

Björn Gambäck
Norwegian University of Science and Technology, Norway

ABSTRACT

Arguably, the most important difference between machines and humans is that humans have feelings. For several decades researchers have been trying to create methods to simulate sentimentality for machines, and currently Sentiment Analysis is the hottest, most demanding, and rapidly growing task in the language processing field. Sentiment analysis or opinion mining refers to the application of Natural Language Processing, Computational Linguistics, and text analytics to identify and extract sentimental (opinionated, emotional) information in a text. The basic task in sentiment analysis is to classify the polarity of a given text at the document, sentence, or feature/aspect level, that is, to decide whether the expressed sentiment in a document, a sentence, or a feature/aspect is positive (happy), negative (sad), neutral (memorable), and so forth. In this chapter, the authors discuss various challenges and solution strategies for Sentiment Analysis with a particular view to texts in Bangla (Bengali).

1. INTRODUCTION: SENTIMENT ANALYSIS ( )

The title of this chapter is inspired by the Bangla science-fiction writer Narayan Sanyal. One of his most popular Sci-Fi novels is *Nakshatraloker Debatatma* [নক্ষত্রলোকেরদেবতাত্মা] (1976), which was inspired by Sir Arthur C. Clarke's novel *2001: A Space Odyssey* (1968). Sanyal's book first describes the evolution of the human race all the way from primitive creatures to intelligent beings building civilisations and ruling the Earth. The book then takes the history further into the space age, with Jupiter exploration and the same

DOI: 10.4018/978-1-4666-3970-6.ch015

super intelligent computer, "HAL" as in Clark's work. Sanyal called HAL "*Jantra-Na*" (যন্ত্রনা), which in Bangla ambiguously means both 'not a machine' (যন্ত্রনা) and 'pain' (যন্ত্রনা), metaphorically portraying the key difference between machines and humans: "The Feelings."

In the late 80s, researchers in Natural Language Processing (NLP) and Artificial Intelligence (AI) started to realize that machines should be able to understand and express sentiment to be intelligent. Since then researchers have attempted textual Sentiment Analysis (SA) for a range of different languages. Sentiment Analysis defines an overall problem, which addresses multiple sub-problems. It is without any doubt a challenging and enigmatic research task. Any scientific research needs to know the proper definitions of its problems in order to solve them. The essential question that is raised at the beginning of the sentiment analysis research is "*What is sentiment or opinion?*" Several researchers have tried to answer this question in the light of a range of research fields, such as Psychology, Philosophy, Psycholinguistics, and Cognitive Science, with many different researchers attempting to give their own definitions, going all the way back to Plato who interpreted opinion as being the medium between Knowledge and Ignorance.

Sentiment analysis research as such started as a content analysis problem in Behavioural Science. The General Inquirer system (Stone et al., 1966) was the first attempt in this direction. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when a document was written. The system usually counted the occurrences of positive or negative emotion instances in any particular piece of text. The General Inquirer system and work by several researchers from the early 90s onwards (e.g., Wiebe et al., 1990; Hatzivassiloglou and McKeown, 1997; Turney, 2002; Pang and Lee, 2004) are milestones that mark the avenues to the current research trends of today. However, although sentiment analysis research started long ago, the question "*What is sentiment or opinion?*" still remains unanswered. It is very hard to define sentiment or opinion, and to identify the regulating or the controlling factors of sentiment. Moreover, it has not been possible to define a concise set of psychological forces that really affect the writers' sentiments, that is, the human sentiment, broadly speaking. Probably the question cannot be answered by the theories of Computer Science, and maybe the scopes of Medicine, Cognitive Science, Psychology, and other science fields have to be explored. Topically Relevant Opinionated Sentiment detection is better known as Subjectivity Detection (Wiebe et al., 1990). Janyce Wiebe borrowed the definition of opinion from Psycholinguistic research such as Quirk et al. (1985) which states that "an opinion could be defined as a private state that is not open to objective observation or verification."

Sentiment Analysis/Opinion Mining from natural language text is thus both a multifaceted and multidisciplinary AI problem (Liu, 2010). It tries to narrow the communication gap between the highly sentimental human and the sentimentally restricted computers by developing computational systems that can recognize and respond to the sentimental states of the human users. There is a perpetual debate about the best ways of collecting intelligence either by following the functional path of biological human intelligence or by generating new methodologies for completely heterogeneous mechatronic machines and defining a completely new horizon called electronic intelligence. Present research endeavors try to find the optimal solution strategies for machines that either mimic the techniques of self-organized biological human intelligence or can at least simulate the functional similarities of human sentimental intelligence.

Though it might even be impossible to formulate a complete analytical definition of sentiment (Kim and Hovy, 2004), the motivation behind the whole sentiment analysis research field is to

develop solution strategies to meet the practical necessities. 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 technologies 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 language processing systems, there is a growing need for sentiment analysis systems that can automatically process the plethora of sentimental information available in online electronic text. The increasing social necessity is the driving force for the massive research efforts on Sentiment Analysis/ Opinion Mining. But why does sentiment analysis become so imperative? Because knowing what others think always is a very important factor in our decision making. For example, before buying electronic products like TVs, laptops, iPads or smartphones, or before going to the cinema to watch the latest *Prometheus* or *Skyfall* we google on it to find out, *What do others think?* about the object or subject. With the proliferation of social networking, a plethora of important information is being added to the World Wide Web every day. Only Twitter adds on average over 400 million messages (tweets) per day. This data offers new and exciting opportunities, and there is much useful information that can be learned from meaningful analysis of the data. It has therefore over the last few years been a growing public and enterprise interest in different types of social media and their role in modern society and especially in sentiment analysis (Burwen 2012; Grimes 2012).

As discussed above, sentiment analysis research first started for the English language, but to satisfy the necessities of multilingual users all over the Globe many researchers have made efforts to develop technologies for other languages. Our endeavor was to develop mechanisms to make machines sensitive to Bangla, or Bengali, as it is known as according to the International Standards Organisation (ISO-639 language code: 'ben'). Bangla is the World's 5th most common language in terms of speakers (over 350 million of which about 200 million have it as first language), the second most common in India and the national language of Bangladesh. The main problem of working with Bangla is the scarcity of electronic resources and the morpho-syntactic richness of the language. When we started back in 2006 there were no resources available for Bangla, and we thus had to develop resources like lexica and corpora, and basic processing tools like a stemmer, part-of-speech tagger, etc., to start the actual research. Those resource creation processes are truly inseparable part of language processing research, especially while working with under-resourced languages such as Bangla. In connection to the difficulties of Bangla in particular, we would like to suggest interested readers to read "Why Indian Languages Failed to Make a Mark Online!" (PJ 2010).

This chapter summarizes the research endeavors by the authors on almost every granular aspects of Sentiment Analysis and especially on Bangla. Bangla is a morpho-syntactically and culturally rich language; therefore sentiment analysis from Bangla is undoubtedly tough in itself. Sentiment is not a direct property of languages; therefore an intelligent system needs some prior knowledge to act senti-mentally. Sentiment knowledge is generally wrapped into a computational lexicon, technically called a Sentiment Lexicon. The development process of such a lexicon for Bengali, the *Bengali SentiWordNet* is described in Section 2 (যন্ত্রানুভূতি-সংকলন). Similar to classical pattern recognition problems, Sentiment Analysis can also be divided into the identification and the classification genres, called *sentiment/subjectivity detection* and *polarity classification*, respectively. The proposed techniques for subjectivity

detection and polarity classification for Bangla are elaborated in Sections 3 (যন্ত্র-বোধদয়) and 4 (যন্ত্রানুভূতি-মেরুধর্মিতা-নিরূপণ).

The needs of the end users are the driving forces behind sentiment analysis research. The end users are often not looking for just 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. Proper *structurization* of sentiments is essential before proceeding to any further granular analysis or generation and aggregation. Structurization involves identification of various aspects of a sentiment/opinion, such as sentiment holder, and sentiment topic. The research attempts on structurization are described in Section 5 (যন্ত্রানুভূতি-পর্যেষণা). To meet the satisfaction level of end users, an intelligent sentimental/opinionated information processing system should be capable of presenting an at-a-glance view of aggregated information, scattered over various sources/documents. Textual or visual summarization, visualization or tracking of sentiment are all striking needs from the perspective of the end users. The overall summarization-visualization-tracking research attempts are described in Section 6 (যন্ত্রঃক্রিয়-অনুভূতি-সাংক্ষেপ). Finally, Section 7 discusses the future of sentiment analysis.

2. BANGLA SENTIWORDNET
(-)

Sentiment knowledge acquisition in terms of a sentiment lexicon is a vital pre-requisite of any sentiment analysis system. Previous studies have proposed to attach *prior polarity* (Esuli and Sebastiani, 2006) to each sentiment lexicon level. Prior polarities are approximate values and are based on corpus statistics. The techniques for the creation of sentiment lexica can broadly be categorized into two types, one follows the classical manual annotation techniques (Andreevskaia and Bergler, 2006; Wiebe and Riloff, 2005; Mohammad et al., 2008) and the other includes various automatic techniques (Tong, 2001; Mohammad and Turney, 2010). Both types of techniques have some limitations. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain. Manual annotation techniques are trustworthy, but in general takes time for development. Manual annotation techniques furthermore require a large number of annotators to balance the sentimentality of individual annotators in order to reach agreement, but qualified human annotators are both costly and difficult to find. There are two issues that should be satisfied by a good quality sentiment lexicon. The first one is coverage and the second is credibility of the associative polarity scores. Automatic processes are good for coverage expansion, but manual methods are trustable for prior polarity assignment. Both the processes have been attempted to develop SentiWordNet(s) (Das and Bandyopadhyay, 2010c; Das and Bandyopadhyay, 2010e) for several languages.

The automatic processes used in the present work are bilingual dictionary based look-up, WordNet-based synonym and antonym expansion, orthographic antonym generation and corpus-based induction. English sentiment lexica were chosen as the source and the synset members were translated into the target language using bilingual dictionaries. WordNet 3.0 was effectively used to expand a given synset via synonym and antonym search. Sixteen hand-crafted suffix/affix rules (like normal – ab-normal, natural – un-natural) were used to orthographically create more antonyms for a given synset, and corpus validation was carried out later to confirm the validity of the orthographically generated forms. The generated sentiment lexicon was used as a seed list. The language specific corpus was automatically tagged with these seed words using the simple tagset of Sentiment Word Positive (SWP) and Sentiment Word Nega-

tive (SWN). A Conditional Random Field (CRF) based classifier was trained on the tagged corpus and then applied to the un-annotated corpus to find out new language and culture specific sentimental words. These techniques have been successfully used for three Indian languages: *Bengali*, *Hindi* and *Telugu* (Das and Bandyopadhyay, 2010c; Das and Bandyopadhyay, 2010e). The Bengali SentiWordNet (Das and Bandyopadhyay, 2010f) has already been made publically available.[1]

As there is a high scarcity of human annotators, it was decided to involve the Internet population for creating more credible sentiment lexica (Das and Bandyopadhyay, 2011; Das, 2011). The Internet population is huge and constantly growing (currently ca 2.4 billion; Miniwatts 2012). It consists of people with various languages, cultures, ages, etc., and thus is not biased towards any particular domain, language or society. An interactive online game called *Dr. Sentiment* was developed to collect players' sentiment by asking a set of simple template-based questions to reveal the sentimental status of the player.[2] The lexica tagged by this system are credible as humans tag them. They are not static sentiment lexica, as the prior polarity scores are updated regularly. On average almost 100 players/day currently play *Dr. Sentiment* throughout the world in different languages. *Global SentiWordNet* (Das and Bandyopadhyay, 2010d), SentiWordNets for 57 languages was developed using Google Translate API.

Dr. Sentiment also helps to capture an overall picture of human social psychology regarding sentiment understanding. The age-wise distribution of players' sentimentality is shown in Figure 1. Sentimentality also changes with gender, as reported in Figure 2, and with the players' geospatial location, as exemplified in Figure 3. There it is shown how the word "blue" has been tagged by different players around the world: surprisingly it has been tagged as positive in one part of the world and negative in another part. Most of the negative tags come from the Middle-East and especially from Islamic countries. This might be based on verse 20:102 of the *Qur'an* in which it says that on the day "the Trumpet is blown" (the Day of Resurrection), the sinners shall be gathered, "blue-eyed" – supposedly with their eyes turning blue with fear, hence giving the word "blue" a bad connotation.

Several types of psychological information are currently being incorporated into the existing

Figure 1. Sentimentality age wise

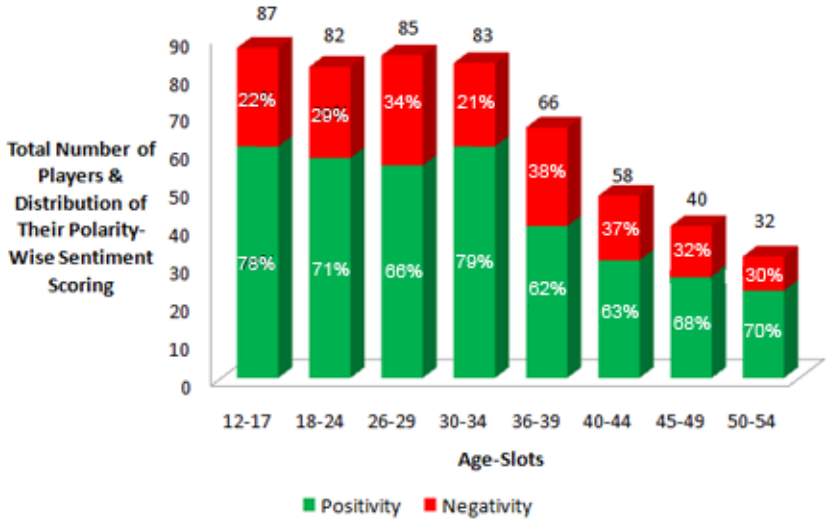
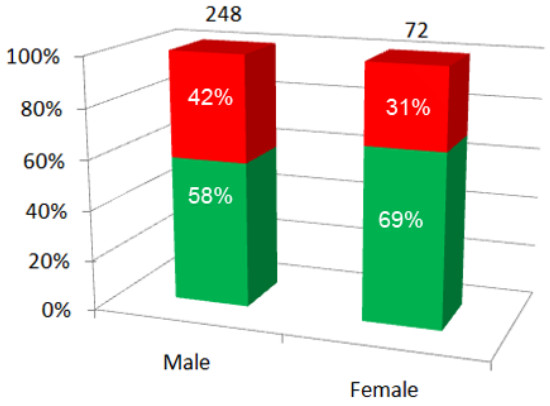


Figure 2. Sentimentality gender wise



SentiWordNet, with the resultant lexicon being termed the *PsychoSentiWordNet* (Das, 2011). The *PsychoSentiWordNet* holds variable prior polarity scores that may be fetched depending upon the regulating psychological aspects. The example in Table 1 illustrates the definition.

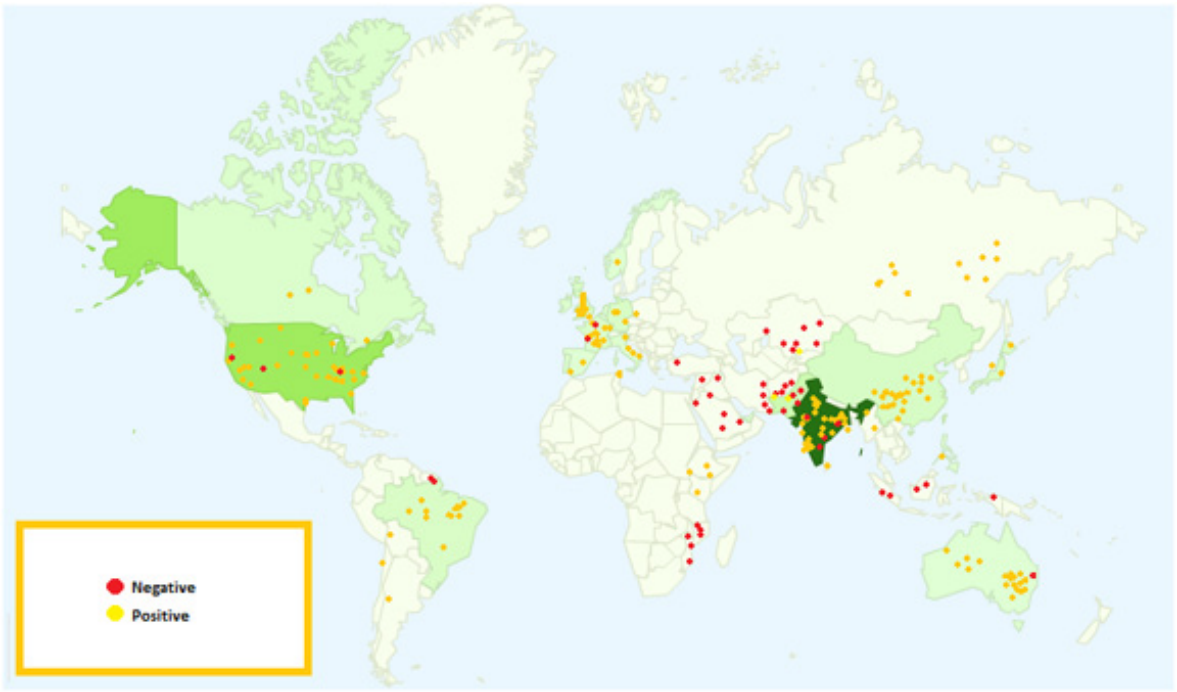
3. SENTIMENT DETECTION AND CLASSIFICATION (-)

The term subjectivity simply refers to the identification of sentiments in a piece of text. More precisely, the term Subjectivity can be defined as the Topically Relevant Opinionated Sentiment (Wiebe *et al*., 1990). The subjectivity is concerned with whether the expressed sentiment is related to the relevant topic or fulfills the overall desired goal of a Sentiment Analysis system.

Table 1. Polarity scores dependent on psychological aspect

Aspect Values (Profession)	Input	Polarity
Null	High	Positive
Businessman	High	Negative
Share Broker	High	Positive

Figure 3. Geospatial sentimentality



Sentiment or subjectivity detection is a very tough challenge for machines with very limited emotional capabilities and even for human beings. Let us take a look at the following examples.

Example 1: Product Review
"My camera broke in two days."

Example 2: Film Review; Film Name: Deep Blue Sea, Holder: Arbitrary-outside of theatre
"This is blue!"

In the first example, it is very hard to disambiguate whether the author is only talking about an accident or complaining about the quality of the camera. The problem with the second example is that there is no evaluative expression and no indicators at syntactic or semantic levels to identify the sentiment. Previous studied have identified some clues at the lexical and syntactic levels (Aue and Gamon, 2005; Hatzivassiloglou and McKeown, 1997; Nasukawa and Yi, 2003). A series of experiments have been carried out to find the optimal feature set for both the English and Bangla languages. The final feature set used for the experiments has been classified into three types (levels) as reported in Table 2.

On the algorithmic aspect, the experiments started with a*rule-based* (Das and Bandyopadhyay, 2009c) technique and continued with Machine Learning (Das and Bandyopadhyay, 2009b) and hybrid techniques (Das and Bandyopadhyay, 2009a). A Theme Detection technique was developed to detect topical relevant sentiments. The themes relate to the topic of any document, but there may be more unrevealed clues based on human psychology or on complex relationships among the linguistic clues for sentiment / subjectivity detection which may not be extracted with present NLP/simple machine learning techniques.

Thus, experiments have been carried out with Genetic Algorithms (Das and Bandyopadhyay, 2010g) to adopt the biological evolutionary path of the human intelligence for machines. The accuracy of the system with the Genetic-Based Machine Learning (GBML) technique reaches 90.22% (MPQA: news) and 93.00% (IMDB: movie review) for English and 87.65% (news) and 90.6% (blog) for Bangla, respectively, as stated in Table 3. Machine learning algorithms when applied to NLP systems generally utilize various combinations of syntactic and semantic linguistic features to identify the most effective feature set. The sentiment/subjectivity detection problem in the present task was viewed as a Multi-Objective or Multi-Criteria Optimization search problem. The experiments started with a large set of possible extractable syntactic, semantic and discourse level features. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. The GBML

Table 2. Features for subjectivity detection

Types	Features
Lexico-Syntactic	Part-of-Speech
	SentiWordNet
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing
Discourse Level	Title of the Document
	First Paragraph
	Average Distribution
	Theme Word

Table 3. Results of the genetic algorithm-based subjectivity classifier

Languages	Domain MPQA	Precision	Recall
English	MPQA	90.22%	96.01%
	IMDB	93.00%	98.55%
Bangla	NEWS	87.65%	89.06%
	BLOG	90.6%	92.40%

technique automatically identifies the best feature set based on the principles of natural selection and survival of the fittest. The identified best fitting feature set is then optimized locally, and global optimization is obtained by a multi-objective optimization technique.

4. SENTIMENT POLARITY DETECTION (- -)

Polarity classification is the classical problem from which Sentiment Analysis started. It involves sentiment/opinion classification into semantic classes such as *positive, negative or neutral* and/or other fine-grained emotional classes like *happy, sad, anger, disgust, surprise,* and maybe others. However, for the present task we stick to standard binary classification, i.e., positive and/or negative. We start by discussing previous research endeavors, in particular elaborating on the birth of prior polarity as a concept, its usage for polarity classification, and the most recent trends in prior polarity research.

Sentiment polarity classification (Is the text positive or negative?) started as a semantic orientation determination problem: by identifying the semantic orientation of adjectives, Hatzivassiloglou *et al*. (1997) proved the effectiveness of empirically building a sentiment lexicon. Turney (2002) suggested positive and negative classification by *Thumbs Up* and *Thumbs Down*, while the concept of a prior polarity lexicon was established with the introduction of *SentiWordNet* (Esuli and Sebastiani, 2006). Higher accuracy for prior polarity identification is very hard to achieve, as prior polarity values are approximations only. Hence, the prior polarity method may not excel alone; additional techniques are required for contextual polarity disambiguation. The use of other NLP or machine learning methods to extend human-produced prior polarity lexica was pioneered by Pang *et al*. (2002). Several researches then tried syntactic-statistical techniques for polarity classification, reporting good accuracy (Seeker *et al*., 2009; Moilanen *et al.*, 2010). With these research efforts the two-step methodology, i.e., sentiment lexicon followed by further NLP techniques, became the standard method for polarity classification.

The existing reported solutions or available systems are still far from perfect or fail to meet the satisfaction level of the end users. The main issue may be that there are many conceptual rules that govern sentiment and there are even more clues (possibly unlimited) that can convey these concepts from realization to verbalization of a human being (Liu, 2010). A recent trend of prior polarity takes a different way for sentiment knowledge representation, following the mental lexicon model to hold the contextual polarity as in human knowledge representation. To this end, Cambria *et al*. (2011) introduced a new paradigm: *Sentic Computing*, in which they use an emotion representation and a Common Sense-based approach to infer affective states from short texts over the Web. Grassi (2009) conceived the Human Emotion Ontology as a high-level ontology supplying the most significant concepts and properties constituting the centerpiece for the description of human emotions. To overcome the problems of the present proximity-based static sentiment lexicon based techniques, we have introduced a new way to represent sentiment knowledge using Vector Space Models. This representation of the sentiment knowledge in the Conceptual Spaces of distributional Semantics will be referred to as Sentimantics. The new models can store dynamic prior polarity with different contextual information (e.g., "*long*", context: *waiting* polarity: -0.25 or "*long*", context: *live* polarity: +0.50). The concept of Sentimantics is clearly an off-spring of the existing prior polarity concept, but we deviate philosophically in terms of contextual dynamicity, and ideologically follow the path of Minsky (2006), Cambria *et al*. (2011) and Grassi

(2009), but with a different notion. The strategy has been tested on both English and Bangla. The intension behind choosing two distinct language families is to establish the credibility of the proposed methods.

Since the two-step methodology is the most common approach to polarity classification in practice, a syntactic-polarity classifier was developed to compare the impact of the proposed Sentimantics concept to the standard polarity classification technique, in order to produce comparative results. Adhering to the standard two-step methodology (i.e., prior polarity lexicon followed by any NLP technique), a Syntactic-Statistical polarity classifier was quickly developed using Support Vector Machines (SVM) with SVMTool (Giménez and Márquez, 2004). The intension behind the development of the syntactic polarity classifier was to examine the effectiveness and the limitations of the standard two-step methodology. The following feature set was used: *Sentiment Lexicon, Negative Words, Stems, Function Words, Part of Speech and Dependency Relations*, as most previous research indicated that these are the prime features to detect the sentimental polarity from text (Das and Bandyopadhyay, 2010h).

The feature ablation, presented in Table 4 proves the accountability of the two-step polarity classification technique. The prior polarity lexicon (completely dictionary-based) approach gives about 50% accuracy; the further improvements of the system are obtained by other NLP techniques.

The entries in a prior polarity lexicon are attached with two probabilistic values, positivity and negativity, but according to the best of our knowledge no previous research clarifies which value to pick in what context – and there is no information about this in the SentiWordNet. The general trend is to pick the highest one, but which the correct one actually is may depend on the context. An example may illustrate the problem better: Suppose the word "high" (Positivity: 0.25, Negativity: 0.125 for "high" from SentiWordNet) is attached with a positive polarity (since its positivity value is higher than its negativity value) in the sentiment lexicon. However, the polarity of the word may vary by its particular use.

- Sensex reaches high^{+}.
- Prices go high^{-}.

Hence further processing is required to disambiguate these types of words. Table 5 shows how many words in the SentiWordNet(s) are ambiguous and need special care. There are 6,619 (English) and 7,654 (Bangla) lexicon entries in SentiWordNet(s) where both the positivity and the negativity values are greater than zero. Similarly, there are 3,187 (English) and 2,677 (Bangla) lexi-

Table 4. Performance of the syntactic polarity classifier by feature ablation

Features	Performance	
	English	Bangla
Sentiment Lexicon	50.50%	47.60%
+Negative Words	55.10%	50.40%
+Stemming	59.30%	56.02%
+ Functional Words	63.10%	58.23%
+ Parts Of Speech	66.56%	61.90%
+Chunk	68.66%	66.80%
+Dependency Relations	76.03%	70.04%

Table 5. Statistics for SentiWordNet (the percentages are based on n/28,430 resp. n/30,000)

Types	English	Bangla
Total number of tokens	115,424	30,000
Positivity>0 OR Negativity>0	28,430	30,000
Positivity>0 AND Negativity>0	**6,619 (23.28%)**	**7,654 (25.51%)**
Positivity>0 AND Negativity=0	10,484 (36.87%)	8,934 (29.78%)
Positivity=0 AND Negativity>0	11,327 (39.84%)	11,780 (39.26%)
Positivity>0 AND Negativity>0 AND |Positivity-Negativity|>=0.2	**3,187 (11.20%)**	**2,677 (8.92%)**

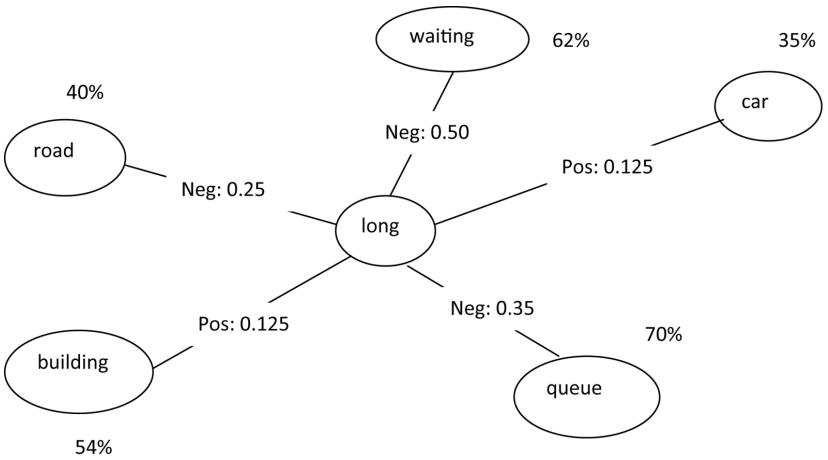
cal entries whose positivity and negativity value difference is less than 0.2. All these lexical entries are ambiguous.

Two different types of models for Sentimantics (Das and Gambäck, 2012) composition have been examined. Both are empirically grounded and can represent the contextual similarity relations among various lexical sentiment and non-sentiment concepts. The experiments started with existing resources such as *ConceptNet* and *SentiWordNet* for English and *SemanticNet* (Das and Bandyopadhyay, 2010n; Das and Bandyopadhyay, 2010p) and *SentiWordNet (Bengali)* for Bangla. The common sense lexica like ConceptNet and SemanticNet were developed for general purposes and the formalization of Sentimantics from these resources faces challenges due to lack of dimensionality. Thus is a second experiment a Vector Space Model (VSM) was developed by a corpus driven semi-supervised method to assign the Sentimantics from scratch. This model performed relatively better than the previous one and was quite satisfactory. Generally, extracting knowledge from this kind of VSM is algorithmically very expensive because the network has a very high dimensionality. An important limitation of this type of model is that it requires very well-defined processed input to extract knowledge such as "Input: (*high*); Context (*sensex, share market, point*)". In the end, a Syntactic Co-Occurrence Based VSM with relatively few dimensions was built. The final model is the best performing lexicon network model and may be described as the acceptable solution to the Sentimantics problem. Each sentiment word in the developed lexical network by the Network overlap technique is assigned a contextual prior polarity. Figure 4 shows the lexical network for the word "long."

5. SENTIMENT STRUCTURIZATION (-)

It is important to keep in mind that the needs of the end users are the driving forces behind the sentiment analysis research: the research endeavors should lead to the development of a real time sentiment analysis system, which successfully satisfies the needs of the end users. Let us have a look at some real life needs of end users. For example, market surveyors from company A may identify the need to find out the changes in public opinion about their product X after release of product Y by another company B. The different aspects of product Y that the public consider better than product X are also points

Figure 4. Sentimantics network developed by the network overlap technique



of interest. These aspects could typically be the durability of the product, power options, weight, color and many more other issues that depend on the particular product. In another scenario, a voter may be interested in studying the change of public opinion about a leader or a public event before and after an election. In this case the aspect could be a social event, economic recession and maybe other issues. The end users are not only looking for binary (positive/negative) sentiment classification, but are more interested in aspectual sentiment analysis. Therefore only sentiment detection and classification is not enough to satisfy the needs of the end users: a sentiment analysis system should be capable of understanding and extracting the aspectual sentiments present in a natural language text.

Previous research efforts have proposed several different structures or components for sentiment extraction. Among the proposed sentiment structures the most widely used structures are *Holder* (Kim and Hovy, 2004; Choi et al., 2005; Bethard et al., 2006), *Topic* (Ku et al., 2005; Zhou et al., 2006; Kawai et al., 2007) and other domain-dependent attributes (Kobayashi et al., 2006; Bal and Saint-Dizier, 2009). However, real life users are not always interested in all the aspects at the same time, but rather look for opinion/sentiment changes of any "Who" during "When" and depending upon "What" or "Where" and the reasons behind "Why." With this hypothesis, we have proposed a 5W (Who/কে, What/কি, When/কখন, Where/কোথায় and Why/কেন) constituent extraction technique for sentiment/opinion structurization (Das et al., 2010i). The 5W structure is domain independent and more generic than the existing semantic constituent extraction structures.

Table 6 presents the sentence level co-occurrence patterns of the 5Ws in the Bangla corpus. The 5Ws do not appear together regularly in the corpus. Hence, sequence labeling with 5W tags using any machine learning technique will lead to a label bias problem and may not be an acceptable solution for the present problem of 5W role labeling. Therefore, a system based on a hybrid architecture has been built. It statistically assigns 5W labels to each chunk in a sentence using Maximum Entropy Modeling (MaxEnt). A rule-based post-processor helps to reduce many false hits by the MaxEnt-based system and at the same time identifies new 5W labels. The rules have been developed based on the acquired statistics on the training set and the linguistic analysis of standard Bangla grammar. By analyzing the output of both the MaxEnt and the hybrid systems (MaxEnt followed by the rule-based post-processor system) it can be easily inferred that the hybrid structure is essential to the 5W problem domain.

6. SENTIMENT SUMMARIZATION (- -)

Aggregation of information is a 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 various types of output formats like textual

Table 6. Sentence level co-occurrence patterns of 5Ws in Bangla

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%

or visual summary, or overall tracking along the time dimension. Several research attempts can be found in the literature on Topic-wise (Yi *et al*., 2003; Pang and Lee, 2004; Zhou *et al*., 2006) and Polarity-wise (Hu, 2004; Yi and Niblack, 2005; Das and Chen, 2007) summarization, and on Visualization (Morinaga *et al*., 2002; Aue and Gamon, 2005; Carenini *et al.*, 2006) and *Tracking* (Lloyd *et al*., 2005; Mishne and de Rijke, 2006; Fukuhara *et al*., 2007). The key issue regarding the sentiment aggregation is how the data shall be aggregated. Dasgupta and Ng (2009) pose an important question: "*Topic-wise, Sentiment-wise, or Otherwise?*" about the opinion summary generation techniques. Actually the output format varies by the end users' requirements and domains. Several output formats have been experimented with in the present work.

The experiments started with multi-document topic-opinion textual summary (Das and Bandyopadhyay, 2010k). A 5W constituent-based textual summarization-visualization-tracking system was devised to meet the need for an at-a-glance presentation. The 5W constituent-based aggregation system is a multi-genre system. The system facilitates users to generate sentiment tracking with a textual summary and a sentiment polarity-wise graph based on any dimension or combination of dimensions they want, for example, "Who" are the actors and "What" 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. The 5W constituent-based summarization-visualization-tracking system aims to cover all genres and attempts to answer the philosophical question "*Topic-Wise*, *Polarity-Wise* or *Other-Wise?*"

- **Topic-Wise:** Users may generate sentiment summaries based on any customized topic like Who, What, When, Where and Why along any dimension or combination of dimensions they want.
- **Polarity-Wise:** The system produces a Gantt 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.

Moreover, the end users can structure their information needs by:

- Who was involved?
- What happened?
- When did it take place?
- Where did it take place?
- Why did it happen?

During the development of the multi-document topic-opinion summarization system, a strong semantic lexical network (Das and Bandyopadhyay, 2010j; Das and Bandyopadhyay, 2010k) was proposed following the idea of Mental Lexicon models. The same lexical semantic network was used to develop the 5W system. The present 5W summarization-visualization-tracking system (Das *et al*., 2012) also provides an overall summary. A snapshot of the 5W Sentiment Summarization-Visualization-Tracking System is presented in Figure 5. Another important aspect of the present system is that a user can leave out the input along a dimension in order to see all the possible information on that dimension.

The working principle of the present 5W summarization-visualization-tracking system is as follows: The system identifies all the desired nodes in the developed semantic constituent network as given by the user in the 5W form. Inter-constituent distances are then calculated from the developed semantic constituent network. For example, suppose the user gave the input shown in Table 7. The calculated inter-constituent distances would then look like those displayed in Table 8.

Next, all the sentences consisting of at least one of the user-defined constituents are extracted from all documents. The extracted sentences are

Figure 5. A snapshot of the 5W summarization-visualization-tracking system

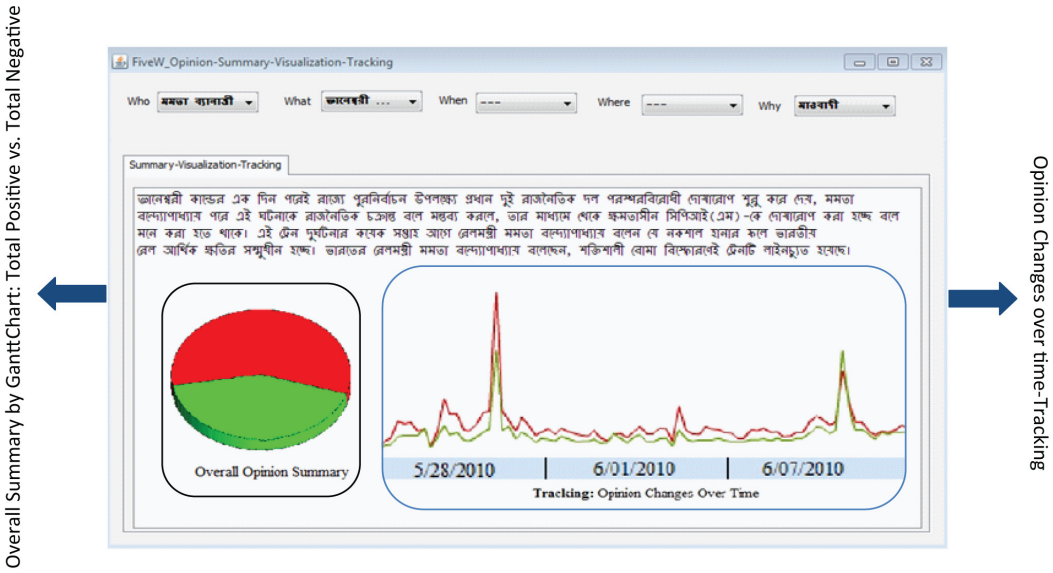


Table 7. Example of 5Ws chosen by end user

	Who	What	When	Where	Why
INPUT	মমতাব্যানার্জী	জ্ঞানেশ্বরীএক্সপ্রেস	মধ্যরাত	ঝাড়গ্রাম	মাওবাদী
	(Mamata Banerjee)	(Gyaneswari Express)	(Midnight)	(Jhargram)	(Maoist)

ranked with the adaptive Information Science Page-Rank algorithm based on the constituents present in the sentence. In the first iteration, the Page-Rank algorithm assigns a score to each sentence based on keyword presence (constituents are treated as keywords at this stage). In the second iteration, the ranks calculated by Page-Rank are multiplied by the inter-constituent distances for those sentences where more than one constituent is present. In the example sentence below two Ws ("Who" and "What") are jointly present. Suppose the assigned rank for the sentence by Page-Rank is *n*. Then in the next iteration the modified score will be *n**0.86, because the inter-constituent distance for "Who" (মমতাবন্দ্যোপাধ্যায়) and "What" (জ্ঞানেশ্বরীএক্সপ্রেস) is 0.86.

Table 8. Calculated inter-constituent distances

Type	Inter-Constituent Distances				
	Who	What	When	Where	Why
Who	-	0.86	0.02	0.34	0.74
What	0.86	-	0.80	0.89	0.67
When	0.02	0.80	-	0.58	0.23
Where	0.34	0.89	0.58	-	0.20
Why	0.74	0.67	0.23	0.20	-

মমতা_বন্দ্যোপাধ্যায়/ **Who** জ্ঞানেশ্বরী_এক্সপ্রেস_ঘটনাকে/ **What** রাজনৈতিকচক্রান্তবলেমন্তব্যকরেন।

English Gloss: Mamta_Bandyopadhyay/ **Who** commented that the Gyaneshwari_Express_incident/ **What** is a political conspiracy.

The ranked sentences are then sorted in descending order and the top-ranked 30% (of all retrieved sentences) are shown as a summary. The ordering of sentences is very important for summarization. We prefer the temporal order of sentences as they occurred in original document, when it was published.

The visual tracking system consists of five drop down boxes. The drop down boxes give options for individual 5W dimension of each unique W that exists in the corpus. An example output from the present 5W summarization-visualization-tracking system is shown in Table 9.

Produced Textual Summary: পরশু মধ্যরাতে ঝাড়গ্রামের অদূরে জ্ঞানেশ্বরী এক্সপ্রেসের লাইনচ্যুত হওয়ার ঘটনাকে বড়সড় রাজনৈতিক ষড়যন্ত্র বলে দাবি করেন মমতা।শ্রীমতী মমতা বন্দ্যোপাধ্যায় পরদিন সকালেই ঝাড়গ্রাম পৌছান ও প্রেসমিটিং-এ জানান, সিবিআইকে দিয়ে ঘটনাটি তদন্ত করা হচ্ছে । তদন্ত শুরু করেছে সিআইডি, তবে রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় ট্রেন বেলাইন হওয়ার কারণ হিসেবে রেল লাইনে বিস্ফোরণ ঘটার তথ্য দিয়েছেন, যার কোনও প্রমাণ পাওয়া যায়নি । এমনকী এই ঘটনা যে পুরভোটের আগে তাঁকে বেকায়দায় ফেলার চক্রান্ত, এমন ইঙ্গিতও দিয়েছেন মমতা বন্দ্যোপাধ্যায়।

English Gloss: Mamta claimed that the derailment incident of the Jyaneswari Express near Jharagramera, which happened at midnight the day before yesterday is a big political conspiracy. Smt. Mamta Bandyopadhyay reached Jharagrama next morning and said in a press conference that the case will be investigated by CBI. CID has started investigation, but rail minister Mamta Bandyopadhyay has presented a theory of an explosion as a probable reason for the derailment of the train, of which no evidence has stillbeen found. This incident before the municipality election is a conspiracy to ensure her defeat, Mamta Bandyopadhyay has indicated.

7. HOW FAR AWAY IS THE "THE BEST-INFORMEDDREAM"OFHALOR ?

Sir Arthur C. Clark's book *2001: A Space Odyssey* was written in 1968 and the ideological replica in Bangla by Narayan Sanyal, *Nakshatraloker Debatatma* [নক্ষত্রলোকেরদেবতাত্মা] in 1976; however, even though approximately four decades have passed after that science fantasy, HAL or "যন্ত্রনা" is still just the "*The Best-Informed Dream*" for researchers in Artificial Intelligence. It is very hard to predict the next probable avenue of this scientific field. Sentiment Analysis is a highly inter-disciplinary research field and it will need to get contributions from research endeavors in disciplines such as Computer Science, AI, Psychology, Philosophy, Psycholinguistics, Cognitive Science, and many more.

How humans understand and express emotions is a complex issue in itself; to make machines understand and express emotions is substantially harder. Textual sentiment analysis is a step in that direction, but it is essential that we manage to identify what particular pieces of text carry what sentiment (at least with some probability). In order to truly start to understand what sentiment and opinion really means, it is also imperative that this is done for many different languages and for people with varying backgrounds and cultures. The research reported in the present chapter is an important contribution in that direction, and

Table 9. Output from the 5W tracking system

	Who	What	When	Where	Why
INPUT	মমতাব্যানার্জী	লাইনচ্যুত	-	ঝাড়গ্রাম	-
	(Mamta Banerjee)	(Derailment)	-	(Jhargram)	-

thus a step towards *understanding* and interpreting sentiment. To be able to do so clearly is a prerequisite for *expressing* emotions. Only when the processes underlying both interpreting and expressing emotions have been fully understood, maybe "the best-informed dream" can be reached and a machine in the future utter, "I am sorry!"

REFERENCES

Andreevskaia, A., & Bergler, S. (2007). Clac and clac-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the 4th SemEval-2007*, ACL.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of RANLP-05*. Borovets, Bulgaria: RANLP.

Bal, K., & Saint-Dizier, P. (2010). Towards building annotated resources for analyzing opinions and argumentation in news editorials. In *Proceedings of LREC 2010*. Valetta, Malta: LREC.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2006). Extracting opinion propositions and opinion holders using syntactic and lexical cues. In The Computing Attitude and Affect in Text: Theory and Applications, (pp. 125-141). Academic Press.

Burwen, M. (2012). *Social media: The end of conventional market research?* Technology Futures.

Cambria, E., Hussain, A., & Eckl, C. (2011). Taking refuge in your personal sentic corner. In *Proceedings of the Workshop on Sentiment Analysis: Where AI meets Psychology*. IJCNLP.

Carenini, G., Ng, R., & Pauls, A. (2006). Multi-document summarization of evaluative text. In *Proceedings of the European Chapter of the Association for Computational Linguistics*. EACL.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the HLT/EMNLP 2005*. HLT/EMNLP.

Das, A. (2010n). Can we mimic human pragmatics knowledge into computational lexicon? In *Proceedings of the International Conference on Natural Language Processing (ICON 2010)*. ICON.

Das, A., & Bandyopadhyay, S. (2009a). Subjectivity detection in English and Bengali: A CRF-based approach. In *Proceeding of the International Conference on Natural Language Processing (ICON 2009)*. ICON.

Das, A., & Bandyopadhyay, S. (2009b). Theme detection an exploration of opinion subjectivity. In *Proceeding of the Affective Computing & Intelligent Interaction (ACII2009)*. Amsterdam, The Netherlands: ACII.

Das, A., & Bandyopadhyay, S. (2009c). Extracting opinion statements from bengali text documents through theme detection. In *Proceeding of the 17th International Conference on Computing (CIC-09)*. GEOS.

Das, A., & Bandyopadhyay, S. (2010c). Dr sentiment creates SentiWordNet(s) for Indian languages involving internet population. In *Proceeding of IndoWordNet Workshop*. ICON.

Das, A., & Bandyopadhyay, S. (2010d). Towards the global SentiWordNet. In *Proceeding of the Workshop on Model and Measurement of Meaning (M3)*. PACLIC.

Das, A., & Bandyopadhyay, S. (2010e). SentiWordNet for Indian languages. In *Proceeding of the 8th Workshop on Asian Language Resources (ALR 8)*. COLING.

Das, A., & Bandyopadhyay, S. (2010f). SentiWordNet for Bangla. In *Proceedings of the Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary (KSE4)*. Mysore, India: KSE.

Das, A., & Bandyopadhyay, S. (2010g). Subjectivity detection using genetic algorithm. In *Proceeding of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA10)*. ECAI.

Das, A., & Bandyopadhyay, S. (2010h). Opinion-polarity identification in Bengali. In *Proceeding of the 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL 2010)*. KESE.

Das, A., & Bandyopadhyay, S. (2010j). Opinion summarization in Bengali: A theme network model. In *Proceeding of the 2nd IEEE International Conference on Social Computing (SocialCom-2010)*. SocialCom.

Das, A., & Bandyopadhyay, S. (2010k). Topic-based Bengali opinion summarization. In *Proceeding of the 23rd International Conference on Computational Linguistics (COLING 2010)*. COLING.

Das, A., & Bandyopadhyay, S. (2010p). SemanticNet-Perception of human pragmatics. In *Proceeding of the 2nd Workshop on Cognitive Aspects of the Lexicon: Enhancing the Structure and Lookup Mechanisms of Electronic Dictionaries (COGALEX-II)*. COLING.

Das, A., & Bandyopadhyay, S. (2011). Dr sentiment knows everything! In *Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Demo Session)*, (pp. 50-55). Portland, OR: ACL.

Das, A., & Gambäck, B. (2012). Sentimantics: The conceptual spaces for lexical sentiment polarity representation with contextuality. In *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*. ACL.

Das, A., Gambäck, B., & Bandyopadhyay, S. (2012). The 5W structure for sentiment summarization-visualization-tracking. In *Proceeding of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*. Delhi, India: CICLING.

Das, A., Ghosh, A., & Bandyopadhyay, S. (2010a). Semantic role labeling for bengali noun using 5Ws: Who, what, when, where and why. In *Proceeding of the International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLPKE2010)*. IEEE.

Das, I. (2011). PsychoSentiWordNet. In *Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Student Session)*. ACL.

Das, S. R., & Chen, M. Y. (2007). Yahoo! For Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388. doi:10.1287/mnsc.1070.0704.

Dasgupta, S., & Ng, V. (2009). Topic-wise, sentiment-wise, or otherwise? Identifying the hidden dimension for unsupervised text classification. In *Proceedings of the EMNLP*. Singapore: EMNLP.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC-2006), (pp. 417-422). Genoa, Italy: LREC.

Fukuhara, T., Nakagawa, H., & Nishida, T. (2007). Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *Proceedings of the International Conference on Weblogs and Social Media.* ICWSM.

Giménez, J., & Márquez, L. (2004). SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: LREC.

Grassi, M. (2009). Developing HEO human emotions ontology. In *Proceedings of the 2009 Joint International Conference on Biometric ID management and Multimodal Communication* (LNCS), (vol. 5707, pp. 244–251). Springer.

Grimes, S. (2012). *DeepMR: Market research mines social sentiment*. GreenBook.

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, (pp. 174–181). Madrid, Spain: ACL.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 168–177). Seattle, WA: ACM.

Kawai, Y., Kumamoto, T., & Tanaka, K. (2007). Fair news reader: Recommending news articles with different sentiments based on user preference. In Proceedings of Knowledge-Based Intelligent Information & Engineering Systems (pp. 612-622). KES.

Kobayashi, N., Iida, R., Inui, K., & Matsumoto, Y. (2006). Opinion mining as extraction of attribute-value relations. *Lecture Notes in Artificial Intelligence*, *4012*, 470–481.

Ku, L.-W., Lee, L.-Y., Wu, T.-H., & Chen, H.-H. (2005). Major topic detection and its application to opinion summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 627-628). Salvador, Brazil: ACM.

Liu, B. (2010). Sentiment analysis: A multi-faceted problem. *IEEE Intelligent Systems*, *25*(3), 76–80.

Lloyd, L., Kechagias, D., & Skiena, S. (2005). Lydia: A system for large-scale news analysis. LNCS. *Proceedings of String Processing and Information Retrieval*, *3772*, 161–166. doi:10.1007/11575832_18.

Minsky, M. (2006). *The emotion machine*. New York: Simon and Schuster.

Mishne, G., & de Rijke, M. (2006). Moodviews: Tools for blog mood analysis. In *Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs* (pp. 153–154). AAAI.

Mohammad, S., Dorr, B., & Hirst, G. (2008). Computing word-pair antonymy. In *Proceedings of the Empirical Methods on Natural Language Processing*. EMNLP.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. NAACL.

Moilanen, K., Pulman, S., & Zhang, Y. (2010). Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. ECAI.

Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). Mining product reputations on the web. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the Conference on Knowledge Capture*. K-CAP.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. ACL.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Empirical Methods on Natural Language Processing*. EMNLP.

PJ. (2010). Why Indian languages failed to make a mark online!. *NextBigWhat*.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Academic Press.

Seeker, W., Bermingham, A., Foster, J., & Hogan, D. (2009). *Exploiting syntax in sentiment polarity classification*. Dublin: Dublin City University.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Boston: The MIT Press.

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussions. In *Proceedings of Working Notes from the Workshop on Operational Text Classification*. ACM.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics*. ACL.

Wiebe, J. (1990). *Recognizing subjective sentences: A computational investigation of narrative text*. (Ph.D. Dissertation). SUNY. Buffalo, NY.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th CICLing-2005*, (pp. 475-486). CICLing.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd ICDM*, (pp. 427-434). Washington, DC: ICDM.

Yi, J., & Niblack, W. (2005). Sentiment mining in WebFountain. In *Proceedings of the International Conference on Data Engineering*. ICDE.

Zhou, L., & Hovy, E. (2006). On the summarization of dynamically introduced information: Online discussions and blogs. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*. Stanford, CA: AAAI.

ENDNOTES

1 http://www.amitavadas.com/sentiwordnet.php

2 http://www.amitavadas.com/Sentiment-Game/