Sentimantics: Conceptual Spaces for Lexical Sentiment Polarity Representation with Contextuality

Amitava Das **Björn Gambäck**
Department of Computer and Information Science
Norwegian University of Science and Technology
Sem Sælands vei 7-9, NO-7094 Trondheim, Norway
amitava.santu@gmail.com gamback@idi.ntnu.no

Abstract

Current sentiment analysis systems rely on static (context independent) sentiment lexica with proximity based fixed-point prior polarities. However, sentiment-orientation changes with context and these lexical resources give no indication of ***which value to pick at what context.*** The general trend is to pick the highest one, but which that is may vary at context. To overcome the problems of the present proximity-based static sentiment lexicon techniques, the paper proposes a new way to represent sentiment knowledge in a Vector Space Model. This model can store dynamic prior polarity with varying contextual information. The representation of the sentiment knowledge in the Conceptual Spaces of distributional Semantics is termed ***Sentimantics.***

1 Introduction

Polarity classification is the classical problem from where the cultivation of Sentiment Analysis (SA) 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 similar. However, for the present task we stick to the standard binary classification, i.e., positive and/or negative.

***The Concept of Prior Polarity*:** Sentiment polarity classification ("*The text is 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 review classification by *Thumbs Up* and *Thumbs Down*, while the concept of prior polarity lexica was firmly established with the introduction of SentiWordNet (Esuli *et al.*, 2004).

More or less all sentiment analysis researchers agree that prior polarity lexica are necessary for polarity classification, and prior polarity lexicon development has been attempted for other languages than English as well, including for Chinese (He *et al.*, 2010), Japanese (Torii *et al.*, 2010), Thai (Haruechaiyasak *et al.*, 2010), and Indian languages (Das and Bandyopadhyay, 2010).

Polarity Classification Using the Lexicon*:** High accuracy for prior polarity identification is very hard to achieve, as prior polarity values are approximations only. Therefore the prior polarity method may not excel alone; additional techniques are required for contextual polarity disambiguation. The use of other NLP methods or machine learning techniques over 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), making the ***two-step methodology (sentiment lexicon followed by further NLP techniques) the standard method for polarity classification.

***Incorporating Human Psychology*:** 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,

Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 38–46,
Jeju, Republic of Korea, 12 July 2012.

2010). The most recent trends in prior polarity adopt an approach to sentiment knowledge representation which lets the mental lexicon model hold the contextual polarity, as in human mental knowledge representation.

Cambria *et al*. (2011) made an important contribution in this direction by introducing a new paradigm: *Sentic Computing*[1], 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.

The Proposed Sentimantics: The present paper introduces the concept of *Sentimantics* which is related to the existing prior polarity concept, but differs from it philosophically in terms of contextual dynamicity. It ideologically follows the path of Minsky (2006), Cambria *et al*. (2011) and (Grassi, 2009), but with a different notion.

Sentiment analysis research started years ago, but still the question "***What is sentiment or opinion?***" remains unanswered! It is very hard to define sentiment or opinion, and to identify the regulating or the controlling factors of sentiment; an analytic definition of opinion might even be impossible (Kim and Hovy, 2004). Moreover, no concise set of psychological forces could be defined that really affect the writers' sentiments, i.e., broadly the human sentiment.

Sentimantics tries to solve the problem with a practical necessity and to overcome the problems of the present proximity-based static sentiment lexicon techniques.

As discussed earlier, the two-step methodology is the most common one in practice. As described in Section 3, a syntactic-polarity classifier was therefore developed, to examine the impact of proposed *Sentimantics* concept, by comparing it to the standard polarity classification technique. The strategy was tested on both English and Bengali. The intension behind choosing two distinct language families is to establish the credibility of the proposed methods.

For English we choose the widely used MPQA[3] corpus, but for the Bengali we had to create our own corpus as discussed in the following section.

The remainder of the paper then concentrates on the problems with using prior polarity values only, in Section 4, while the Sentimantics concept proper is discussed in Section 5. Finally, some initial conclusions are presented in Section 6.

2 Bengali Corpus

News text can be divided into two main types: (1) news reports that aim to objectively present factual information, and (2) opinionated articles that clearly present authors' and readers' views, evaluation or judgment about some specific events or persons (and appear in sections such as 'Editorial', 'Forum' and 'Letters to the editor'). A Bengali news corpus has been acquired for the present task, based on 100 documents from the 'Reader's opinion' section ('Letters to the Editor') from the web archive of a popular Bengali newspaper.[4] In total, the corpus contains 2,235 sentences (28,805 word forms, of which 3,435 are distinct). The corpus has been annotated with positive and negative phrase polarities using Sanchay[5], the standard annotation tool for Indian languages. The annotation was done semi-automatically: a module marked the sentiment words from SentiWordNet (Bengali)[6] and then the corpus was corrected manually.

3 The Syntactic Polarity Classifier

Adhering to the standard two-step methodology (i.e., prior polarity lexicon followed by any NLP technique), a Syntactic-Statistical polarity classifier based on Support Vector Machines (SVMs) has been quickly developed using SVMTool.[7] The intension behind the development of this syntactic polarity classifier was to examine the effectiveness and the limitations of the standard two-step methodology at the same time.

The selection of an appropriate feature set is crucial when working with Machine Learning techniques such as SVM. We decided on a feature

[1] http://sentic.net/sentics/

[3] http://www.cs.pitt.edu/mpqa/
[4] http://www.anandabazar.com/
[5] http://ltrc.iiit.ac.in/nlpai_contest07/Sanchay/
[6] http://www.amitavadas.com/sentiwordnet.php
[7] http://www.lsi.upc.edu/~nlp/SVMTool/

Polarity	Precision		Recall	
	Eng.	Bng.	Eng.	Bng.
Total	76.03%	70.04%	65.8%	63.02%
Positive	58.6%	56.59%	54.0%	52.89%
Negative	76.3%	75.57%	69.4%	65.87%

Table 1: Overall and class-wise results of syntactic polarity classification

set including *Sentiment Lexicon, Negative Words, Stems, Function Words, Part of Speech and Dependency Relations*, as most previous research agree that these are the prime features to detect the sentimental polarity from text (see, e.g., Pang and Lee, 2005; Seeker et al., 2009; Moilanen et al., 2010; Liu et. al., 2005).

Sentiment Lexicon: SentiWordNet 3.0 [8] for English and SentiWordNet (Bengali) for Bengali.

Negative Words: Manually created. Contains 80 entries collected semi-automatically from both the MPQA[9] corpus and the Movie Review dataset[10] by Cornell for English. 50 negative words were collected manually for Bengali.

Stems: The Porter Stemmer[11] for English. The Bengali Shallow Parser[12] was used to extract root words (from morphological analysis output).

Function Words: Collected from the web. [13] Only personal pronouns are dropped for the present task. A list of 253 entries was collected manually from the Bengali corpus.

POS, Chunking and Dependency Relations: The Stanford Dependency parser[14] for English. The Bengali Shallow Parser was used to extract POS, chunks and dependency relations.

The results of SVM-based syntactic classification for English and Bengali are presented in Table 1, both in total and for each polarity class separately.

To understand the effects of various features on the performance of the system, we used the feature ablation method. The dictionary-based approach using only SentiWordNet gave a 50.50% precision

Features	Precision	
	Eng.	Bng.
Sentiment Lexicon	50.50%	47.60%
+Negative Words	55.10%	50.40%
+Stemming	59.30%	56.02%
+ Function Words	63.10%	58.23%
+ Part of Speech	66.56%	61.90%
+Chunking	68.66%	66.80%
+Dependency Relations	76.03%	70.04%

Table 2: Performance of the syntactic polarity classifier by feature ablation

(Eng.) and 47.60% (Bng.) which can be considered as baselines. As seen in Table 2, incremental use of other features like negative words, function words, part of speech, chunks and tools like stemming improved the precision of the system to 68.66% (Eng.) and 66.80% (Bng.). Further use of syntactic features in terms of dependency relations improved the system precision to 76.03% (Eng.) and 70.04% (Bng.). The feature ablation proves the accountability of the two-step polarity classification technique. The prior polarity lexicon (completely dictionary-based) approach gives about 50% precision; the further improvements of the system are obtained by other NLP techniques.

To support our argumentation for choosing SVM, we tested the same classification problem with another machine learning technique, Conditional Random Fields (CRF)[15] with the same data and setup. The performance of the CRF-based model is much worse than the SVM, with a precision of 70.04% and recall of 67.02% for English, resp. 61.23% precision and 55.00% recall for Bengali. The feature ablation method was also tested for the CRF model and the performance was more or less the same when the dictionary features and lexical features were used (i.e., SentiWordNet + Negative Words + Stemming + Function Words + Part of Speech). But it was difficult to increase the performance level for the CRF by using syntactic features like chunking and dependency relations. SVMs work excellent to normalize this dynamic situation.

It has previously been noticed that multi-engine based methods work well for this type of heterogeneous tagging task, e.g., in Named Entity

[8] http://sentiwordnet.isti.cnr.it/
[9] http://www.cs.pitt.edu/mpqa/
[10] http://www.cs.cornell.edu/People/pabo/movie-review-data/
[11] http://tartarus.org/martin/PorterStemmer/java.txt
[12] ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php
[13] http://www.flesl.net/Vocabulary/Single-word_Lists/function_word_list.php
[14] http://nlp.stanford.edu/software/lex-parser.shtml
[15] http://crfpp.googlecode.com/svn/trunk/doc/index.html

Recognition (Ekbal and Bandyopadhyay, 2010) and POS tagging (Shulamit et al., 2010). We have not tested with that kind of setup, but rather looked at the problem from a different perspective, questioning the basics: *Is the two-step methodology for the classification task ideal or should we look for other alternatives?*

4 What Knowledge at What Level?

In this section we address some limitations regarding the usage of prior polarity values from existing of prior polarity lexical resources. Dealing with unknown/new words is a common problem. It becomes more difficult for sentiment analysis because it is very hard to find out any contextual clue to predict the sentimental orientation of any unknown/new word. There is another problem: word sense disambiguation, which is indeed a significant subtask when applying a resource like SentiWordNet (Cem *et al*., 2011).

A prior polarity lexicon is 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 SentiWordNet. The general trend is to pick the highest one, but which may vary by context. An example may illustrate the problem better: Suppose a word "*high*" (Positivity: 0.25, Negativity: 0.125 from SentiWordNet) is attached with a positive polarity (its positivity value is higher than its negativity value) in the sentiment lexicon, but the polarity of the word may vary in any particular use.

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

Hence further processing is required to disambiguate these types of words. Table 3 shows how many words in the SentiWordNet(s) are ambiguous and need special care. There are 6,619 (Eng.) and 7,654 (Bng.) lexicon entries in SentiWordNet(s) where both the positivity and the negativity values are greater than zero. Therefore these entries are ambiguous because there is no clue in the SentiWordNet which value to pick in what context. Similarly, there are 3,187 (Eng.) and 2,677 (Bng.) lexical entries in SentiWordNet(s) whose positivity and negativity value difference is less than 0.2. These are also ambiguous words.

Types	Eng.	Bng.
	Numbers (%) English: n/28,430 Bengali: n/30,000	
Total Token	115,424	30,000
Positivity > 0 ∨ Negativity > 0	28,430	30,000
Positivity > 0 ∧ Negativity > 0	6619 (23.28 %)	7,654 (25.51 %)
Positivity > 0 ∧ Negativity = 0	**10,484 (36.87 %)**	**8,934 (29.78 %)**
Positivity = 0 ∧ Negativity > 0	**11,327 (39.84 %)**	**11,780 (39.26 %)**
Positivity > 0 ∧ Negativity > 0 ∧ |Positivity-Negativity| ≥ 0.2	3,187 (11.20 %)	2,677 (8.92 %)

Table 3: SentiWordNet(s) statistics

The main concern of the present task is the ambiguous entries from SentiWordNet(s). The basic hypothesis is that if we can add some sort of contextual information with the prior polarity scores in the sentiment lexicon, the updated rich lexicon network will serve better than the existing one, and reduce or even remove the need for further processing to disambiguate the contextual polarity. How much contextual information would be needed and how this knowledge should be represented could be a perpetual debate. To answer these questions we introduce ***Sentimantics****: Distributed Semantic Lexical Models to hold the sentiment knowledge with context.*

5 Technical Solutions for Sentimantics

In order to propose a model of Sentimantics we started with existing resources such as ConceptNet [16] (Havasi *et al*., 2007) and SentiWordNet for English, and SemanticNet (Das and Bandyopadhyay, 2010) and SentiWordNet (Bengali) for Bengali. The common sense lexica like ConceptNet and SemanticNet are developed for general purposes, and to formalize Sentimantics from these resources is problematic due to lack of dimensionality. Section 5.1 presents a more rational explanation with empirical results.

In the end we developed a Syntactic Co-Occurrence Based Vector Space Model to hold the Sentimantics from scratch by a corpus driven semi-supervised method (Section 5.2). This model performs better than the previous one and quite satisfactory. Generally extracting knowledge from

[16] http://csc.media.mit.edu/conceptnet

this kind of VSM is very expensive algorithmically because it is a very high dimensional network. Another important limitation of this type of model is that it demands very well defined processed input to extract knowledge, e.g., ***Input:*** *(high)* ***Context:*** *(sensex, share market, point*). Philosophically, the motivation of Sentimantics is to provide a rich lexicon network which will serve better than the existing one and reduce the requirement of further language processing techniques to disambiguate the contextual polarity. This model consists of relatively fewer dimensions. The final model is the best performing lexicon network model, which could be described as the acceptable solution for the Sentimantics problem. The details of the proposed models are described in the following.

5.1 Semantic Network Overlap, SNO

We started experimentation with network overlap techniques. The network overlap technique finds overlaps of nodes between two lexical networks: namely ConceptNet-SentiWordNet for English and SemanticNet-SentiWordNet (Bengali) for Bengali. The working principle of the network overlap technique is very simple. The algorithm starts with any SentiWordNet node and finds its closest neighbours from the commonsense networks (ConceptNet or SemanticNet). If, for example, a node chosen from SentiWordNet is "long/লম্বা", the closest neighbours of this concept extracted from the commonsense networks are: "road (40%) / waiting (62%) / car (35%) / building (54%) / queue (70%) …" The association scores (as the previous example) are also extracted to understand the semantic similarity association. Hence the desired *Sentimantics* lexical network is developed by this network overlap technique. The next prime challenge is to assign contextual polarity to each association. For this a corpus-based method was used; based on the MPQA[17] corpus for English and the corpus developed by us for. The corpora are pre-processed with dependency relations and stemming using the same parsers and stemmers as in Section 3. The dependency relations are necessary to understand the relations between the evaluative expression and other modifier-modified chunks in any subjective sentence. Stemming is necessary to understand the root form of any word and for dictionary comparison. The corpus-driven method assigns each sentiment word in the developed lexical network a contextual prior polarity, as shown in Figure 1.

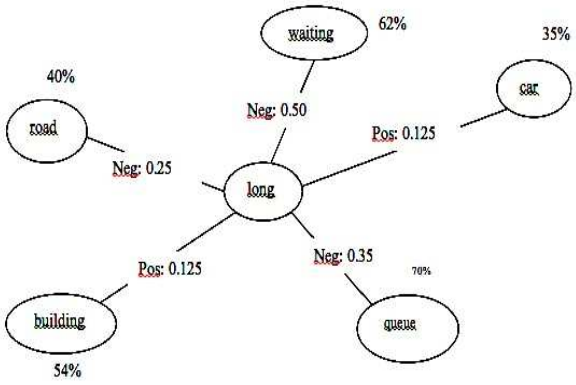

Figure 1: The Sentimantics Network

Semantic network-based polarity calculation

Once the desired lexical semantic network to hold the Sentimantics has been developed, we look further to leverage the developed knowledge for the polarity classification task. The methodology of contextual polarity extraction from the network is very simple, and only a dependency parser and stemmer are required. For example, consider the following sentence.

`We have been waiting in a long queue.`

To extract the contextual polarity from this sentence it must be known that *waiting-long-queue* are interconnected with dependency relations, and stemming is a necessary pre-processing step for dictionary matching. To extract contextual polarity from the developed network the desired input is (*long*) with its context (*waiting, queue*). The accumulated contextual polarity will be Neg: (0.50+0.35)=0.85. For comparison if the score was extracted from SentiWordNet (English) it would be Pos: 0.25 as this is higher than the negative score (*long*: Pos: 0.25, Neg: 0.125 in SentiWordNet).

SNO performance and limitations

An evaluation proves that the present Network Overlap technique outperforms the previous syntactic polarity classification technique. The precision scores for this technique are 62.3% for English and 59.7% for Bengali on the MPQA and

[17] http://www.cs.pitt.edu/mpqa/

Type	Number		Solved By Semantic Overlap Technique
Positivity > 0 ∧ Negativity > 0	Eng.	6,619	2,304 (34.80 %)
	Bng.	7,654	2,450 (32 %)
|Positivity - Negativity| ≥ 0.2	Eng.	3,187	957 (30 %)
	Bng.	2,677	830 (31.5 %)

Table 4: Results of Semantic Overlap

Bengali corpora: clearly higher than the baselines based on SentiWordNet (50.5 and 47.6%; Table 2).

Still, the overall goal to "*reduce/remove the requirement to use further NLP techniques to disambiguate the contextual polarity*" could not be established empirically. To understand why, we performed an analysis of the errors and missed cases of the semantic network overlap technique: most of the errors were caused by lack of coverage. ConceptNet and SemanticNet were both developed from the news domain and for a different task. The comparative coverage of SentiWordNet (English) and MPQA is 74%, i.e., if we make a complete set of sentiment words from MPQA then altogether 74% of that set is covered by SentiWordNet, which is very good and an acceptable coverage. For Bengali the comparative coverage is 72%, which is also very good. However, the comparative coverage of SentiWordNet (English)-ConceptNet and SentiWordNet (Bengali)-SemanticNet is very low: 54% and 50% respectively: only half of the sentiment words in the SentiWordNets are covered by ConceptNet (Eng) resp. SemanticNet (Bng).

Now look at the evaluation in Table 4 which we report to support our empirical reasoning behind the question "*What knowledge to keep at what level?*" It shows how much fixed point-based static prior polarity is being resolved by the Semantic Network Overlap technique. The comparative results are noteworthy but not satisfactory: only 34% (Eng.) and 32% (Bng.) of the cases of "*Positivity* > 0 ∧ *Negativity* > 0" resp. 30% (Eng.) and 31.5 % (Bng.) of the cases of "|*Positivity* - *Negativity*| ≥ 0.2" are resolved by this technique. The results are presented in Table 4.

As a result of the error analysis, we instead decided to develop a Vector Space Model from scratch in order to solve the Sentimantics problem and to reach a satisfactory level of coverage. The experiments in this direction are reported below.

5.2 Starting from Scratch: Syntactic Co-Occurrence Network Construction

A syntactic word co-occurrence network was constructed for only the sentimental words from the corpora. The syntactic network is defined in a way similar to previous work such the Spin Model (Takamura *et al*., 2005) and Latent Semantic Analysis to compute the association strength with seed words (Turney and Litman, 2003). The hypothesis is that all the words occurring in the syntactic territory tend to have similar semantic orientation. In order to reduce dimensionality when constructing the network, only the open word classes *noun*, *verb*, *adjective* and *adverb* are included, as those classes tend to have maximized sentiment properties. Involving fewer features generates VSMs with fewer dimensions.

For the network creation we again started with SentiWordNet 3.0 to mark the sentiment words in the MPQA corpus. As the MPQA corpus is marked at expression level, SentiWordNet was used to mark only the lexical entries of the subjective expressions in the corpus. As before, the Stanford POS tagger and the Porter Stemmer were used to get POS classes and stems of the English terms, while SentiWordNet (Bengali), the Bengali corpus and the Bengali processors were used for Bengali.

Features were extracted from a ±4 word window around the target terms. To normalize the extracted words from the corpus we used CF-IOF, concept frequency-inverse opinion frequency (Cambria *et al*., 2011), while a Spectral Clustering technique (Dasgupta and Ng, 2009) was used for the in-depth analysis of word co-occurrence patterns and their relationships at discourse level. The clustering algorithm partitions a set of lexica into a finite number of groups or clusters in terms of their syntactic co-occurrence relatedness.

Numerical weights were assigned to the words and then the cosine similarity measure was used to calculate vector similarity:

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

When the lexicon collection is relatively static, it makes sense to normalize the vectors once and store them, rather than include the normalization in the similarity metric (as in Equation 2).

$$s\left(\vec{q}_k, \vec{d}_j\right) = \frac{\sum_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} \times \sqrt{\sum_{j=1}^{N} w_{j,k}^2}} \quad \text{-------(2)}$$

ID	Lexicon	1	2	3
1	Broker	**0.63**	0.12	0.04
1	NASDAQ	**0.58**	0.11	0.06
1	**Sensex**	**0.58**	0.12	0.03
1	*High*	**0.55**	0.14	0.08
2	India	0.11	**0.59**	0.02
2	**Population**	0.15	**0.55**	0.01
2	*High*	0.12	**0.66**	0.01
3	Market	0.13	0.05	**0.58**
3	**Petroleum**	0.05	0.01	**0.86**
3	UAE	0.12	0.04	**0.65**
3	*High*	0.03	0.01	**0.93**

Table 5: Five example cluster centroids

After calculating the similarity measures and using a predefined threshold value (experimentally set to 0.5), the lexica are classified using a standard spectral clustering technique: Starting from a set of initial cluster centers, each document is assigned to the cluster whose center is closest to the document. After all documents have been assigned, the center of each cluster is recomputed as the centroid or mean $\vec{\mu}_j$ (where $\vec{\mu}_j$ is the clustering coefficient) of its members:

$$\vec{\mu} = \left(1/\left|c_j\right|\right)\sum_{x \in c_j} \vec{x}$$

Table 5 gives an example of cluster centroids by spectral clustering. Bold words in the lexicon name column are cluster centers. Comparing two members of $Cluster_2$, '**India**' and '**Population**', it can be seen that '**India**' is strongly associated with $Cluster_2$ (p=0.59), but has some affinity with the other clusters as well (e.g., p=0.11 with $Cluster_1$). These non-zero values are still useful for calculating vertex weights during the contextual polarity calculation.

Polarity Calculation using the Syntactic Co-Occurrence Network

The relevance of the semantic lexicon nodes was computed by summing up the edge scores of those edges connecting a node with other nodes in the same cluster. As the cluster centers also are interconnected with weighted vertices, inter-cluster relations could be calculated in terms of weighted network distance between two nodes within two separate clusters.

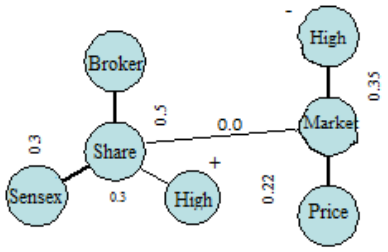

Figure 2: Semantic affinity graph for contextual prior polarity

As an example, the lexicon level semantic orientation from Figure 2 could be calculated as follows:

$$S_d(w_i, w_j) = \frac{\sum_{k=0}^{n} v_k}{k} * w_j^p \quad \text{----(3) or}$$

$$= \sum_{c=0}^{m} \frac{\sum_{k=0}^{n} v_k}{k} * \prod_{c=0}^{m} l_c * w_j^p \text{---(4)}$$

Where $S_d(w_i,w_j)$ is the semantic orientation of w_i with w_j given as context. Equations (3) and (4) are for intra-cluster and inter-cluster semantic distance measure respectively. k is the number of weighted vertices between two lexica w_i and w_j. v_k the weighted vertex between two lexica, m the number of cluster centers between them, l_c the distance between their cluster centers, and w^p_j the polarity of the known word w_j.

This network was created and used in particular to handle unknown words. For the prediction of semantic orientation of an unknown word, a bag-of-words method was adopted: the bag-of-words chain was formed with most of the known words, syntactically co-located.

A classifier based on Conditional Random Fields was then trained on the corpus with a small set of features: co-occurrence distance, ConceptNet similarity scores, known or unknown based on SentiWordNet. With the help of these very simple features, the CRF classifier identifies the most probable bag-of-words to predict the semantic orientation of an unknown word. As an example: Suppose *X* marks the unknown words and that the probable bag-of-words are:

```
9_11-X-Pentagon-USA-Bush
Discuss-Terrorism-X-President
Middle_East-X-Osama
```

Once the target bag-of-words has been identified, the following equation can be used to calculate the polarity of the unknown word *X*.

Discuss-0.012-Terrorism-0.0-X-0.23-President

The scores are extracted from ConceptNet and the equation is:

$$w_x^p = \sum_{i=0}^{n} e_i * \sum_{j=1}^{n} p_i \quad \text{-----(5)}$$

Where e_i is the edge distances extracted from ConceptNet and P_i is the polarity information of the lexicon in the bag-of-words.

The syntactic co-occurrence network gives reasonable performance increment over the normal linear sentiment lexicon and the Semantic Network Overlap technique, but it has some limitations: it is difficult to formulate a good equation to calculate semantic orientation within the network. The formulation we use produced a less distinguishing value for different bag of words. As example in Figure 2:

$$(\textbf{High}, \textit{Sensex}) = \frac{0.3+0.3}{2} = 0.3$$

$$(\textit{Price}, \textbf{High}) = \frac{0.22+0.35}{2} = 0.29$$

The main problem is that it is nearly impossible to predict polarity for an unknown word. Standard polarity classifiers generally degrade in performance in the presence of unknown words, but the Syntactic Co-Occurrence Network is very good at handling unknown or new words.

The performance of the syntactic co-occurrence measure on the corpora is shown in Table 6, with a 70.0% performance for English and 68.0% for Bengali; a good increment over the Semantic Network Overlap technique: about 45% (Eng.) and 41% (Bng.) of the "*Positivity* > 0 ∧ *Negativity* > 0" cases and 43% (Eng.) and 38% (Bng.) of the "|*Positivity* – *Negativity*| ≥ 0.2" cases were resolved by the Syntactic co-occurrence based technique.

To better aid our understanding of the developed lexical network to hold Sentimantics we visualized this network using the Fruchterman Reingold force directed graph layout algorithm (Fruchterman and Reingold, 1991) and the NodeXL [18] network analysis tool (Smith et al., 2009).

Type	Number		Solved By Syntactic Co-Occurrence Network
Positivity>0 && Negativity>0	**Eng.**	6,619	2978 (45 %)
	Bng.	7,654	3138 (41 %)
|Positivity-Negativity|>=0.2	**Eng.**	3,187	1370 (43 %)
	Bng.	2,677	1017 (38 %)

Table 6: Results of the syntactic co-occurrence based technique

6 Conclusions

The paper has introduced *Sentimantics*, a new way to represent sentiment knowledge in the Conceptual Spaces of distributional Semantics by using in a Vector Space Model. This model can store dynamic prior polarity with varying contextual information. It is clear from the experiments presented that developing the Vector Space Model from scratch is the best solution to solving the Sentimantics problem and to reach a satisfactory level of coverage. Although it could not be claimed that the two issues "*What knowledge to keep at what level?*" and "*reduce/remove the requirement of using further NLP techniques to disambiguate the contextual polarity*" were fully solved, our experiments show that a proper treatment of Sentimantics can radically increase sentiment analysis performance. As we showed by the syntactic classification technique the lexicon model only provides 50% accuracy and further NLP techniques increase it to 70%, whereas by the VSM based technique it reaches 70% accuracy while utilizing fewer language processing resources and techniques.

To the best of our knowledge this is the first research endeavor which enlightens the necessity of using the dynamic prior polarity with context. It is an ongoing task and presently we are exploring its possible applications to multiple domains and languages. The term *Sentimantics* may or may not remain in spotlight with time, but we do believe that this is high time to move on for the dynamic prior polarity lexica.

[18] http://www.codeplex.com/NodeXL

References

Cambria Erik, Amir Hussain and Chris Eckl. 2011. Taking Refuge in Your Personal Sentic Corner. SAAIP, IJCNLP, pp. 35-43.

Cem Akkaya, Janyce Wiebe, Conrad Alexander and Mihalcea Rada. 2011. Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis. CoNLL.

Das Amitava and Bandyopadhyay S. 2010. SemanticNet-Perception of Human Pragmatics. COGALEX-II, COLING, pp 2-11.

Das Amitava Bandyopadhyay S. 2010. SentiWordNet for Indian Languages. ALR, COLING, pp 56-63.

Dasgupta, Sajib and Vincent Ng. 2009. Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification. EMNLP.

Ekbal A. and Bandyopadhyay S. 2010. Voted NER System using Appropriate Unlabeled Data. *Lingvisticae Investigationes Journal*.

Esuli Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. LREC, pp. 417-422.

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

Grassi, Marco. 2009. Developing HEO Human Emotions Ontology. Joint International Conference on Biometric ID management and Multimodal Communication, vol. 5707 of LNCS, pp 244–251.

Haruechaiyasak Choochart, Alisa Kongthon, Palingoon Pornpimon and Sangkeettrakarn Chatchawal. 2010. Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews. ALR, pp 64–71.

Hatzivassiloglou Vasileios and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, pp. 174–181.

Havasi, C., Speer, R., Alonso, J. 2007. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. RANLP.

He Yulan, Alani Harith and Zhou Deyu. 2010. Exploring English Lexicon Knowledge for Chinese Sentiment Analysis. CIPS-SIGHAN, pp 28-29.

Kim Soo-Min and Eduard Hovy. 2004. Determining the Sentiment of Opinions. COLING, pp. 1367-1373.

Liu Bing. 2010. *NLP Handbook*. Chapter: Sentiment Analysis and Subjectivity, 2nd Edition.

Liu Hugo, Henry Lieberman and Ted Selker. 2003. A Model of Textual Affect Sensing using Real-World Knowledge. IUI, pp. 125-132.

Minsky Marvin. 2006. *The Emotion Machine*. Simon and Schuster, New York.

Moilanen Karo, Pulman Stephen and Zhang Yue. 2010. Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression. WASSA, pp. 36--43.

Ohana Bruno and Brendan Tierney. 2009. Sentiment classification of reviews using SentiWordNet. In the 9th IT&T Conference.

Pang Bo, Lillian Lee and Vaithyanathan Shivakumar. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP, pp 79-86.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. ACL, pp. 115-124.

Seeker Wolfgang, Adam Bermingham, Jennifer Foster and Deirdre Hogan. 2009. Exploiting Syntax in Sentiment Polarity Classification. National Centre for Language Technology Dublin City University, Ireland.

Shulamit Umansky-Pesin, Roi Reichart and Ari Rappoport. 2010. A Multi-Domain Web-Based Algorithm for POS Tagging of Unknown Words. COLING.

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. 4th International Conference on Communities and Technologies, pp. 255-264.

Takamura Hiroya, Inui Takashi and Okumura Manabu. 2005. Extracting Semantic Orientations of Words using Spin Model. ACL, pp. 133-140.

Torii Yoshimitsu, Das Dipankar, Bandyopadhyay Sivaji and Okumura Manabu. 2011. Developing Japanese WordNet Affect for Analyzing Emotions. WASSA, ACL, pp. 80-86

Turney Peter and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Turney Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ACL, pp. 417–424.

Turney Peter. 2006. Similarity of Semantic Relations. *Computational Linguistics*, 32(3):379-416.