# Literature Survey

A lot of research has taken place in Sentiment Analysis in the past decade. As discussed in chapter 1, bag-of-words strategy poses various problems in analysing the sentiments of the opinions or reviews. Because of that most of the research has been focussed on developing sophisticated supervised approaches for Sentiment Classification. In this chapter we will discuss works done in various sub-fields of sentiment analysis. We start with some of the sophisticated supervised and unsupervised approaches to sentiment classification in Section 3.1. In Section 3.2, we talk about various methods of feature selection. Section 3.3 describes work on language discourse coherent features followed by approaches to handle conditional sentences in Section 0. Finally, Section 3.5 presents different approaches in sentiment analysis for the language Hindi and various works on Discourse on Hindi language.

## 3.1 Approaches to Sentiment Classification

Sentiment classification is the task of classifying the document into various categories like positive, negative or neutral. We can generalise a sentiment classification problem as, "Given an opinionated piece of text, wherein it is assumed that the overall opinion in it is about one single issue or item, we need to classify the opinion as falling under one of two opposing sentiment polarities."

Some of the earliest works in Sentiment Classification were done by Turney (2002) and Pang and Lee (2002). While Turney (2002) came up with an unsupervised approach defining the semantic orientation of words to aid sentiment based classification of reviews, Pang and Lee (2002) implemented a bunch of supervised approaches to sentiment classification in order to maximize the accuracies obtained. In this section, we'll discuss these approaches and a few more similar approaches to Sentiment Classification.

### 3.1.1 An Unsupervised Approach to Sentiment Classification

Turney P (2002) devised an unsupervised learning algorithm that can classify any given review as either *thumbs up* or *thumbs down*. The algorithm can be divided into three steps:

a. Identify phrases in the input text that contain adjectives or adverbs. The assumption here is that the modifier words like adjectives and adverbs are the most important sentiment bearing words.
b. The next step is to estimate the *Semantic Orientation* (SO) of each extracted phrase. Semantic Orientation is calculated using the PMI-IR algorithm. Both Semantic Orientation and PMI-IR algorithm are described later.
c. The final step is to assign the review to one of the classes, recommended or not recommended, based on the average semantic orientation of the phrases extracted from the review.

**Semantic Orientation**: The semantic orientation of any phrase is said to be positive, if it has good association (e.g. romantic ambience) and it is said to be negative, if the associations are bad (e.g. horrific events).

Semantic Orientation is calculated using the Point-wise Mutual Information (PMI) and Information Retrieval (IR), which give a similarity of pairs of words or phrases.

$$SO\ (phrase)\ =\ PMI\ (phrase, excellent)\ -\ PMI\ (phrase, poor)$$

Turney P (2002) experimented on reviews on eight different topics from four varied domains. The accuracies ranged from 65.83% on reviews in movie domain to 84.00% on reviews in automobile domain.

### 3.1.2 Supervised and Semi-Supervised Approaches to Sentiment Classification

Pang and Lee (2002) carried out one of the first set of experiments in Sentiment Analysis using basic supervised approaches, taking a leaf out of standard text based categorization. They implemented Naive Bayes, MaxEnt as well as an SVM-based classifier on a simple bag-of-words feature set for sentiment classification. They tested these systems on a Movie-review dataset generated from IMDB. With these experiments, they received a best case accuracy of 82.9% in case of unigram features with SVM classifier. During this work, they also try out bigram features and POS-specific features, however, SVMs provided the best results. This work has more or less served as the baseline for all future works in supervised sentiment classification.

Pang and Lee (2004) came up with a more sophisticated approach that was based on the intuition that only the subjective portions of the text actually affect the polarity of the document. Hence, they implemented a min-cut based subjectivity classifier before actual sentiment analysis could take place. The subjectivity classifier produces a min-cut based classification using a Naive Bayesian classifier to separate the subjective portions from the objective portions of the text. The subjectivity extract (all subjective sentences from the subjectivity classifier) were then passed on to a general polarity classifier (typically SVM) to decide the overall polarity of the text. The basic flow of the algorithm is described in Figure 0.1

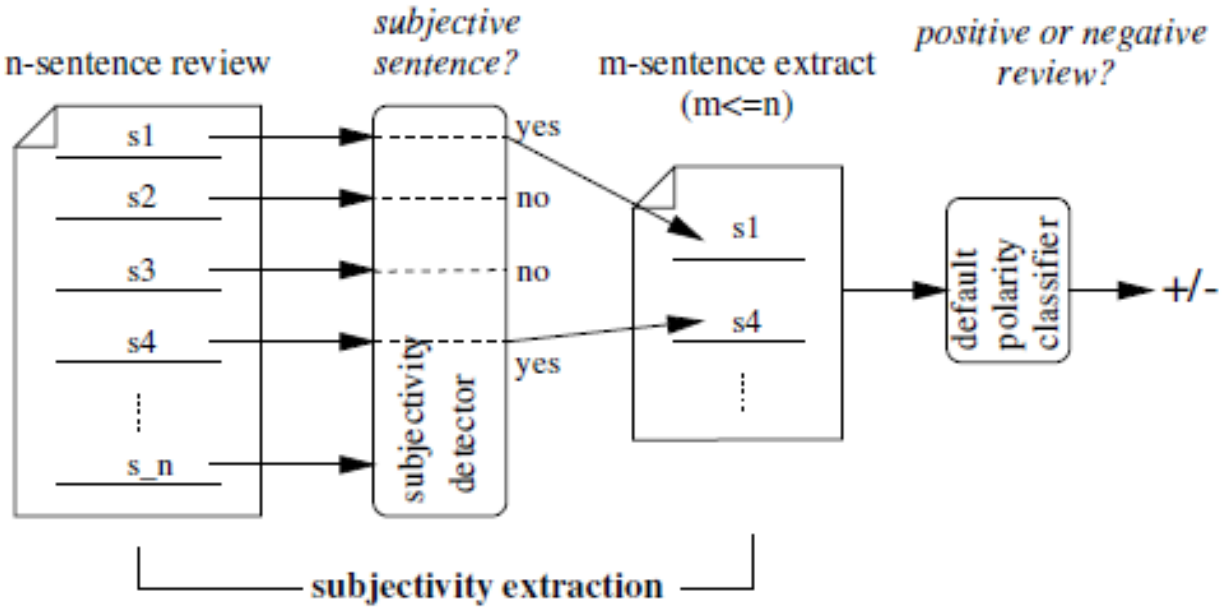


*Figure 0.1: Flow chart for min-cut based subjectivity detection for sentiment classification* [5]

Using this approach, Pang and Lee (2004) showed that taking the subjectivity extracts out and applying sentiment analysis to it alone gives better accuracies than the baseline, *i.e.*, applying to the complete text. They achieved an accuracy of around 86%, which is an improvement of 4% over the baseline. In their experiments, they also showed that a subjectivity extract that produces this accuracy contains only 60% of the original text size. Thus, this approach benefits on two fronts, namely, improving the accuracy and at the same time, summarizing the sentiment content of the review.

Both these approaches tend to be naive as they focus on using simple bag-of-words based features for sentiment classification. Instead, Matsumoto et al. (2005) discussed the usage of syntactic relations between words in the same sentence for sentiment classification. The underlying philosophy is that other than n-grams (a general bag-of-words), word order and syntactic relations between words, too, are intuitively important in sentiment classification. They used word subsequences and dependency sub-trees to represent word order and syntactic relationships respectively. The results show an increase of around 5% in the accuracy over the baseline (bag-of-words).

Socher et al. (2011) introduced a novel machine learning framework based on recursive autoencoders that does sentence level prediction of sentiment label distributions. They tried to learn vector space representations for multiword phrases. They showed that these representations outperform other state-of-the art approaches on commonly used datasets such as movie reviews, without using any sentiment lexicons or polarity shifting rules. They also introduced a new dataset that contain distributions over a broad range of human emotions. This dataset was a part of an experience project which consists of personal user stories annotated with multiple labels which, when aggregated, forms a multinomial distribution that captures emotional reactions.

Christopher Potts was an instructor of a Sentiment Analysis Symposium tutorial **Ошибка! Источник ссылки не найден.**which aimed to cover all aspects of building effective sentiment analysis systems for textual data, with and without sentiment relevant metadata like star ratings. Initially, they focused on imporant aspects of sentiment analysis which needs to be kept in mind while building systems for sentiment analysis like sentiment is blended and multidimensional, sentiment is context dependent etc. Then, they shifted their focus on text preparation tools like *Tokenization, Stemming, Negation, Part-of-speech and dependency parsing.*

They also discussed various lexicons like *Bing Liu's Opinion lexicon, MPQA Subjectivity lexicon, SentiWordNet, general Inquirer and LIWC* and also discussed relationships between them. They also learned about how to build our own lexicons. Then was the turn of various classifiers and vector space models for sentiment analyses like *Naïve Bayes, maximum Entropy etc.*, various feature selection models and comparisons between them. They concluded the tutorial with sentiment summarization. They created tools and demos for the academia which are publicly available on the website.

The approaches mentioned above give a snapshot of some of the common approaches that have been explored in sentiment analysis. A lot more modifications and enhancements have happened since, however, a discussion on those is out of scope of this report. In the following sections, we will discuss the literature specifically relevant to our work during this project.

## 3.2 Approaches to Feature Selection

Document Classification is a problem of automatically assigning pre-defined labels to the documents. We can think sentiment classification as a special kind of document classification where labels are positive and negative. As more and more textual information is available on the internet, effective selection of text, important for accurate sentiment classification has becomes an extremely complex task. Accuracy of document classification is greatly affected by the number and quality of features used. If a model containing low quality features is used, accuracy is greatly affected.

With lots of data to process, it becomes difficult to deal with. Feature Selection is also a method of dimensionality reduction, *i.e.* removal of non informative terms without the loss of classification accuracy. It is an important problem of selecting the relevant features of a document.

Most of the initial research in the field of sentiment classification has been bag-of-words feature representation. Approaches using only bag-of-words feature representation does not provide great results and can serve only as the baseline of further improvements.
In this work, we have discussed various feature selection methods and their importance for sentiment classification.

### 3.2.1 Various Feature Selection Methods

Yang and Pedersen (1997) discussed a Comparative Study of 5 Feature Selection method for Effective Document Categorization.

**a) Document Frequency (DF):**

It is the number of documents in which a term occurs. It's the simplest approach in which those terms are removed whose DF is less than a predefined number. Its complexity is linear with the number of documents. It removes rare terms which are non-informative. Thus, it would be better if rare terms consist of noise.

**b) Information Gain :**

It is the number of bits of information by knowing the presence or absence of a term in a document.

$$G(t) = -\sum_{i=1}^{m} \Pr(c_i)\, log\, \Pr(c_i) + Pr(c_i) \sum_{i=1}^{m} \Pr(c_i|t)\, log\, \Pr(c_i|t)$$

Its Complexity is O(Vm),
Where
V: Vocabulary Size
m: m-ary classification

**c) Mutual Information (MI):**

It is the criteria usually used in statistical language modelling of word associations. If one considers a following contingency table:

| | Term t Occurs | Term t does not occur |
|---|---|---|
| **Category c occur** | A | B |
| **Category c does not occur** | C | D |

*Table 0.1: Contingency Table*

Then,

$$MI(t,c) = \log \frac{\Pr(t \wedge c)}{\Pr(t) * \Pr(c)}$$

It is estimated using

$$MI(t,c) = \log \frac{A*N}{(A+C)*(A+B)}$$

Where N: total number of documents.

It has a value zero if the term and the category are independent. For a term, it is calculated as

$$MI(t) = \sum_{i=1}^{m} \Pr(c_i)\, I(t,c)$$

It has a complexity of O(Vm). A weakness of this approach is that the score is strongly influenced by the marginal probabilities of terms. Thus, for terms with equal conditional probability$\Pr(t|c)$, rare terms have a higher score than the common terms. Thus, the scores are not comparable across terms of widely differing frequency.

**d) Chi–Squared Statistic (CHI):**

It is the lack of independence between t and c. It's defined as

$$CHI^2(t,c) = \frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)}$$

It has a value zero if the term and the category are independent. For a term, it is calculated as

$$CHI^2(t) = \sum_{i=1}^{m} \Pr(c_i)\, CHI^2(t,c)$$

It has a complexity of O(Vm) similar to MI and IG. It is a normalized value and hence is comparable across terms of widely differing frequency. But normalization breaks down if any of the term in the contingency table is tightly populated as in the case of rare terms. Hence, it is not reliable for low frequency terms.

**e) Term Strength (TS):**

It is based on how commonly a term is likely to appear in closely related documents. Let x and y be an arbitrary pair of distinct but related documents and t be a term. Then, it is defined as

$$s(t) = Pr(t \in y \mid t \in x)$$

It is based on document clustering, assuming that documents with many shared words are related and those terms in heavily overlapping area are relatively important.

Yang and Pedersen (1997) found that CHI and IG performs the best. Also there is a strong correlation between DF and IG and DF and CHI which suggests that one can use low cost DF instead of CHI and IG.

Simeon and Hilderman (2008) talked about a new feature selection method Categorical Proportional Difference. It compares CPD with 5 other feature selection methods: CHI, TS, DF, MI and odds ratio.

**f) Categorical Proportional Difference (CPD):**

CPD is the measure of the degree to which a word contributes to differentiating a particular category from other categories.CPD of a word in a particular category is the ratio of a number of documents in a category to the number of documents from other categories where same word occurs.

$$CPD(w,c) = \frac{\#Documen\ ts\ in\ a\ category\ \ where\ w\ occurs}{\#Documents\ \ in\ other\ categories\ \ where\ w\ occurs}$$

Or,

$$CPD(w,c) = \frac{A-B}{A+B}$$

**g) Odds Ratio (OR):**

It measures the odds of a word occurring in the positive class normalized by that of negative class.

$$OR(w,c) = \frac{AD}{BC}$$

Simeon and Hilderman (2008) showed that CPD performs better most of the time than other feature selection methods.

It would be interesting to use combinations of CPD, DF, CHI and IG to find the relevant features for sentiment analysis.

## 3.3 Language and Discourse Coherent Features on Sentiment Classification in English

In Rhetorical Structure Theory, Soricut and Marcu (2003) developed probabilistic models (generative) for identifying elementary discourse units at clausal level and generating trees at the sentence level using lexical and syntactic information from discourse-annotated corpus of RST. But this work ignored the sequential and hierarchical dependencies between the constituents in the parsing model. Also, they made an independence assumption between the label and the structure while modeling a constituent.

Polanyi and Zaenen (2004) described how lexical and discourse items affects valence of an item. They investigated the effects of intensifiers, negatives, modals and connectors that changes the prior polarity or valence of the words and brings out a new meaning or perspective. They also talked about pre-suppositional items and irony and present a simple weighting scheme to deal with them. The authors looked at the effect of multi-entity evaluation and the

genre or attitude assessment of the speaker. They also presented their analysis on reported speech and sub-topics.

Wolf and Gibson (2005) presented a set of discourse structure relations and a way to code or represent them. They reported a method for annotating discourse coherent structures and found different kinds of crossed dependencies. They also indicated statistically that tree cannot be a data structure for storing discourse structure relations.

Wellner et al. (2006) considered the problem of automatically identifying arguments of discourse connectives in the PDTB. They modeled the problem as a predicate-argument identification where the predicates were discourse connectives and arguments served as anchors for discourse segments.

Narayanan, Liu and Choudhary (2009) performed an extensive study of the effect of conditional sentences in Sentiment Analysis. They found that around 8% sentences in a typical document consist of conditional sentences which are difficult to analyze from the point of view of SA due to the conditional clauses. The authors build a supervised learning model to determine if the sentiment expressed by the conditional statements is positive, negative or objective. They used words as features with their POS tags, position in the head or body of the clause, tense pattern, length, negation words and some other syntactic features.

Recasens et al. (2013) proposed an approach to predict the *lifespan of a discourse* with the help of syntactic and semantic features. All discourse constituents are not alive for the same period. Some appear in a variety of discourse contexts (coreferent), whereas others dying out after just one mention (singletons). This constitutes the lifespan of a discourse. The author builds a logistic regression model for predicting the singleton/coreferent distinction taking linguistic insights from syntactic and semantic features.

The features used were *internal morphosyntax of the mention* that concern the internal morphology and syntactic structure of the mention, *grammatical role of the mention* and *semantic environment of the mention*. They have shown the effectiveness of their model. They also showed that incorporating it into a state-of-the-art coreference resolution system yields a significant improvement. This distinction of singleton/coreferent would benefit not only co-reference resolution, but also topic analysis, textual entailment, and discourse coherence.

Alec et al. (2009) provided one of the first studies on sentiment analysis on micro-blogging websites. Barbosa et al. (2010) and Bermingham et al. (2010) both cited noisy data as one of the biggest hurdles in analyzing text in such media. Alec et al. (2009) described a distant supervision-based approach for sentiment classification. They used hashtags in tweets to create training data and implemented a multi-class classifier with topic-dependent clusters.

Barbosa et al. (2010) proposed an approach to sentiment analysis in Twitter using POS-tagged n-gram features and some Twitter specific features like hashtags. Joshi et al., 2011 described a rule based sentiment analysis system for tweets, C-Feel-It. It uses *inquirer*, *sentiwordnet*, *subjectivity lexicon* and *taboada* as lexicons for finding the word polarity. It follows the bag-of-words approach.

Most of the previous works focus on extracting discourse segments or connectives from the text and identifying their span. We view the problem from the angle of the effect these discourse elements have in analyzing the sentiment in the document. All discourse elements are

not essential for SA. Some of them enhance the sentiment, some express hypothetical situations, some reverse the sentiment polarity and some of them express cause-effect relations. We give more importance to those linguistic constructs that either reinforce sentiment or reverse them and ignore those that express irrealis events.

Our work on discourse coherent features (discussed in chapter 0) builds on the work of Polanyi and Zaenen (2004), Wolf and Gibson (2005) and Wellner et al. (2006) and carries the idea further.

Also our system is inspired from C-Feel-IT, a Twitter based sentiment analysis system (Joshi et al., 2011). However, our system is an enhanced version of their rule based system with specialized modules to tackle Twitter spam, text normalization and entity specific sentiment analysis.

## 3.4 Approaches to handle Conditional Statements

Handling conditional sentences is a tricky problem as these sentences are not always true in the real world. They are called irrealis sentences. There are four basic types of conditional sentences (Wren and Martin; 1973-2010 and Narayanan et al.; 2009): Zero conditional, First Conditional, Second Conditional and Third Conditional. Carmen Diaconescu, Valahia University of Târgovişte explained various special cases of conditional sentences apart from the usual four types.

There has been not much work to handle conditional statements for the purpose of sentiment analysis. Narayanan et al. (2009) proposed a supervised technique to handle conditional sentences. They used various features like Sentiment words/phrases and their locations, part-of-speech tags of sentiment words, objective words, tense patterns, special characters like *'!'* and *'?'*, conditional connectives, length of the conditional and consequent clause and negation words. They performed experiments at clause as well as sentence based level.

Our work on handling conditional sentences is different than the approach followed by Narayanan et al. *(2009*). We have used linguistic knowledge to convert these Irrealis sentences into sentences in the real world. We have performed experiments on the Twitter Dataset as well as the dataset used in Narayanan et al. *(2009*).

## 3.5 Language and Discourse Coherent Features in Hindi Sentiment Classification

Prasad et al. (2009) presented their initial efforts towards developing a large scale corpus of Hindi texts annotated with discourse relations. They adopted the lexically grounded approach of the Penn Discourse Treebank (PDTB) and presented a preliminary analysis of discourse connectives in a small corpus. They described how discourse connectives are represented in the sentence-level dependency annotation in Hindi, and discussed how the discourse annotation can enrich this level for research and applications. The main goal of their work was to build a Hindi Discourse Relation Bank along the lines of the PDTB.

Oza et al. (2009) described the Hindi Discourse Relation Bank project. This was in continuation of the work done by Prasad et al. (2009) where they adopted the lexically grounded approach of the Penn Discourse Treebank. They described the classification of Hindi discourse connectives, modifications to the sense classification of discourse relations, and some cross-linguistic comparisons based on some initial annotations carried out so far. They described three types of discourse relations: *explicit connectives*, which are "closed class" expressions drawn from well-defined grammatical classes, *alternative lexicalizations* (AltLex), which are non-connective expressions that cannot be defined as explicit connectives and *implicit connectives*, which are implicit discourse relations "inferred" between adjacent sentences not related by an explicit connective.

Along the same line of work on Hindi Discourse relation Bank project, Kolachina et al. (2012) described experiments on evaluating the discourse relation annotation scheme of the Penn Discourse Treebank (PDTB), in the context of annotating discourse relations in Hindi

Discourse Relation Bank (HDRB). They indicated that overall, some of the changes has made the annotation task much more difficult for the annotators, as also reflected in lower inter-annotator agreement for the relevant sub-tasks. They emphasized the importance of best practices in annotation task design and guidelines.

There has been very little work in the area of sentiment analysis on an Indian language. Joshi et al. (2010) proposed a fall-back strategy to do sentiment analysis for Hindi documents. They studied three approaches to perform SA in Hindi. The first of their approaches involved training a classifier on this annotated Hindi corpus and using it to classify a new Hindi document. In the second approach, they translated the given document in Hindi into English and used a classifier trained on Standard English movie reviews to classify the document. In the third approach, they developed a lexical resource called Hindi-SentiWordNet (H-SWN) and implemented a majority score based strategy to classify the given document. They showed that best results will be achieved with an annotated corpus in the same language of analysis. They also developed a sentiment annotated corpora in the Hindi movie review domain.

Our work is inspired from Joshi et al. (2010). We are going to use linguistic knowledge in the form of discourse and use a simple bag of words model as a baseline system.

## 3.6 Sentiment Analysis Research at IITB

IIT Bombay has been very active in the field of Sentiment Analysis for around half a decade now. Throughout this time, all our efforts have mainly focused on generalizing various aspects of Sentiment Analysis. Agarwal and Bhattacharyya (2005) discussed an approach wherein they incorporated linguistic knowledge in Sentiment Analysis using WordNet Synonymy Graph. Joshi et al (2011) and Balamurali et al took Sentiment Analysis to a new language, from lexeme space to a new feature space (sense space). This work on Harnessing Discourse Features takes a leaf from Joshi et al (2011) in that we try to incorporate more linguistic knowledge in Sentiment Analysis. We also follow the idea developed in Balamurali et al (2011) by using WordNet synsets instead of words in Sentiment Classification.

Ramteke et al. (2013) proposed an approach to detect thwarting in product review domain. They define thwarting as "*Thwarting is looked upon as the phenomenon of polarity reversal at a higher level of ontology compared to the polarity expressed at the lower level*". They used a rule based approach based on ontology as their baseline. And they showed that machine learning with annotated corpora is more effective than the rule based approach. This is first of a kind attempt to tackle thwarting. Some of the features used in their approach are Document polarity, number of flips of sign (*i.e.* change of polarity from positive to negative and vice versa), the maximum and minimum values in a sequence, the length of the longest contiguous subsequence of positive values (LCSP), the length of the longest contiguous subsequence of negative values (LCSN) *etc*.