CHAPTER 2

REVIEW OF LITERATURE

2.1 INTRODUCTION

The rich sources of short text available in the web need to be analysed for various purposes like query analysis, classifying web search results, classifying or grouping blog posts and tweets, opinion mining etc. Unlike long documents there are several additional issues that have to be dealt with short text analysis. First, feature selection and reduction is the most important step in text classification. But due to the data sparseness feature of short texts this has became a challenge. Moreover semantic similarity is hard to achieve due to the lack of content and context. Second, the description of short text is concise and need not have regular grammar. Due to this, standard NLP techniques cannot achieve desired results. Third, short text analysis should address the volume of information, dynamic nature of the internet and the availability of training data. This chapter reviews important works carried out in this area. Figure 2.1 shows the general flow diagram used to review the literature. The flow diagram illustrates the way in which how we narrow down the short text analysis problem in web to the specific problem of identifying same wavelength communities from large scale tweets. First section discusses few important works which carried out to solve the problems in short text analysis in general. Homophily is the basic principle behind any OSN. How different kinds of activities and parameters are used in literature to study the human behaviour is discussed in the next section. Opinion is identified as one of the important dimension which induces
homophily. Different approaches to sentiment analysis and their correlation with user behaviour is discussed in the next section. Different ways of identifying communities are discussed in the last section.

![Flow diagram for the review of literature](image)

**Figure 2.1 Flow diagram for the review of literature**

### 2.2 SHORT TEXT ANALYSIS IN WEB

Most of the prior work on short text analysis has focused on data sparseness problem. After fetching the text from its source, which could be a database repository, blog, a file system or the World Wide Web, should be pre-processed for feature extraction. Then short text similarity measures, classification or clustering, sentiment detection, etc. can be applied. This section discusses some of the major works carried out to solve the short text analysis problem in general.
One of the intuitive methods that can be applied to eliminate data sparseness problem is to extend the sparse features of short text with additional information to make it appear like a long text or document. Phan et al (2008) not only uses the explicit user defined categories in Wikipedia but extracts hidden topic of Wikipedia articles to gain more knowledge. They have introduced a new method to classify short text snippets based on Latent Dirichlet Allocation (LDA) (David et al 2003). One of the important problems of enhancing the feature set by using external knowledge is the “Curse of Dimensionality”. Sriram et al (2010) proposed a small set of domain-specific features extracted from the authors profile and text to classify tweets to a predefined set of generic classes such as news, events, opinions, deals and private messages. Wei et al (2009) proposed association rule for short text feature extension and Faguo et al (2010) suggested rule and statistics to feature extraction. Liu et al (2010) used feature selection model based on parts of speech and HowNet (knowledge base for feature selection) for blog mining.

As the amount of short text snippets generated in the web is very huge, users often face the problem of information overload. Clustering could be one of the solutions to this problem. Short text clustering is considered to be complex due to the low frequencies of vocabulary terms in short texts. Clustering will be harder if the domain is narrow (vocabulary overlapping level of the short documents is very high) (Pinto 2008, Pinto et al 2011).

Banerjee et al (2007) used titles of selected Wikipedia articles to augment the text quality and clustered short text snippets based on the enriched representation. They used a snapshot of the Wikipedia database and all queries were directed to this database upon which the Lucene index was built. Hui et al (2007) used N-gram feature extraction and RPCL (Rival Penalized Competitive Learning) (Xu et al 1993) to cluster Chinese short
texts for mining web topics based on Chinese chunks. RPCL is a clustering algorithm used commonly for speech recognition and image segmentation clustering. Hu et al (2009) enhanced existing features by using both WordNet and Wikipedia. They exploited internal semantics to provide a deep understanding of the original short texts and external semantics which incorporate the concepts derived from the world knowledge to reduce the semantic gap. They also proposed a novel hierarchical resolution phase to parse through the short text and categorize them into segments, phrases and words. From this pool, seed phrases were carefully selected which formed the queries to Wikipedia and WordNet.

More recent work by Ni et al (2011) proposed a new short text clustering strategy, Termcut. They modelled collection of short text snippets as a graph $G = (V,E)$, where a vertex represents a text snippet and each weighted edge between two vertices measures the relationship between the two vertices. The core terms are found on the basis of minimizing a new criterion, $RMcut$. The $RMcut$ criterion is a clustering quality criterion, which measures the quality of clusters according to the clustering principle ‘minimizing the intercluster similarity while maximizing the intra-cluster similarity’. Pinto et al (2011) introduced novel measures to automatically determine whether corpus is made up of narrow domain short texts or not. They proposed a domain-independent self-term expansion methodology to enrich baseline corpus by adding co-related terms from an automatically constructed lexical knowledge resource obtained from the same target data set (and not from an external resource). They used this technique to cluster scientific abstracts which belongs to narrow domain.

Extending the short text with external sources like web search may not be good for all type of problems. For instance, the volume of tweets generated is very huge. Using external sources to analyse the sentiments is not
considered as a feasible solution. Instead several approaches to Twitter sentiment analysis (Speriosu et al 2011, Tan et al 2011, Hu et al 2013) used linkage and social information exist in the Twitter OSN to analyse the Twitter sentiments. Related works on Twitter sentiment analysis are discussed in the coming sections.

2.3 HOMOPHILY IN ONLINE SOCIAL NETWORKS

All Online Social Networks (OSNs) follow the principle of homophily: similarity breeds connection (Mcpherson et al 2001). Social scientists have studied extensively the socio-demographic, behavioural and interpersonal characteristics. They used the traditional mode of collecting the data through online, offline and mixed-mode surveys. But recently, the rich data from various OSNs has attracted significant attention from the research community.

Some of the previous work primarily focused on usage statistics and sequences of user activities in OSNs to analyse user behaviour. Benevenuto & Rodrigues (2009) used click-stream data to capture behaviour of OSN users. They provided a click stream model and observed that silent interactions like profile browsing dominate other visible activities. Guo et al (2009) analysed users posting behaviour of original content and observed that 20% users contribute 80% total content in the network. Jiang et al (2010) also analysed the latent and visible interactions in OSN. They conducted a study on Renren, a largest OSN available in China. They constructed a latent interaction graph to capture browsing activity among OSN users. They also observed that latent interactions dominate visible interactions. Lewis et al (2008) created a facebook dataset and they analysed how socio-demographic dimensions like gender, race and ethnicity are correlated with certain network activities. A recent work (Moore, K & McElroy 2012) examined the role of five dimensions (Openness, Conscientiousness, Extroversion, Agreeableness,
Neuroticism) of personality on facebook usage and features. They observed that certain personality traits are correlated with facebook usage.

A recent work (Panigrahy et al 2012) examined how position in the network, activities and user preferences are correlated. They provided a new affinity measure based on distance and conducted studies on email graph and Twitter mention graph. They identified the homophily in terms of demography, queries and tweets among the closely connected users.

Users in the OSNs can have multiple affiliations or dimensions. Analysing multiple social dimensions of users exposed to social network environment is known as collective behavioural analysis (Tang & Liu 2010b). Tang et al (2012) provided a novel framework to extract latent social dimensions based on network connectivity and constructed a discriminative classifier to determine relevant social dimensions. Behavioural prediction can be made from the learned data model. A recent work (Sachan et al 2012) used topics, social graph topology and nature of user interactions to discover latent communities in social graphs.

Some of the recent works (Abbasi et al 2012, Tan et al 2011) also considered connected users in the Twitter domain to study the behavioural correlation. Tan et al (2011) observed that the probability of sharing the same opinion is high if they are connected. Abbasi et al (2012) have selected an online community which resembles a real world community in terms of race, language, religion etc. They extracted tweets related with Arab Spring to analyse the mood before and after the event. They observed that Yemenis were more concerned about security whereas Egyptians were more concerned about revolution and freedom.
2.4 SENTIMENT ANALYSIS IN SOCIAL WEB

Sentiment analysis or opinion mining is a branch of Natural Language Processing (NLP) which focuses mainly on polarity detection and emotion recognition from various kinds of texts. Traditional sentiment analysis was mainly on structured documents, reviews etc. The advent of Web 2.0 and the large volume of user generated content in the form of short texts make the sentiment analysis challenging. Sentiment analysis require a deep understanding of the explicit and implicit, regular and irregular, and syntactical and semantic language rules (Cambria et al 2013).

Sentiment analysis on social web has been utilized for several applications like product reviews (eg. Turney (2002)), movie reviews (eg. Pang & Lee 2005), Analysing political opinions (eg. Golbeck et al 2010), predicting stock market (eg. Bollen et al 2011). In addition, generation of tweets on diverse issues invited researchers to think about domain independent solutions to solve problems like discovering latent communities based on sentiments (Sachan et al 2012), real world behaviour analysis (Abbasi et al 2012) etc.

Several works (Pang & Lee 2008, Huifeng et al 2009, Tsytsarau & Palpanas 2011, Mostafa 2013,) has discussed in detail the sentiment detection in general. Major works on sentiment analysis have been done on product review and movie reviews. Different approaches have been used to analyse sentiments (Prabowo & Thelwall 2010). Pattern and lexicon based approaches (eg. Yi et al 2003, Hiroshi et al 2004, Turney 2002, Joshi et al 2011), Machine learning approaches (eg. Pang et al 2002, Pang & Lee 2005), Hybrid approaches (eg. Prabowo & Thelwall 2010). Accuracy of machine learning approaches mostly depends on the training data. Training data (labelled data) is not available for all domains. Hence the classifier which
works well on one set of data need not perform well for another set of data. Lexicon based approaches depends on the size of sentiment lexicons.

The following section discusses the major recent works carried out in the Twitter sentiment analysis.

2.4.1 Sentiment Analysis in Twitter

Linguistic flexibility in writing tweets poses additional challenges in determining polarity of the tweets. Limited number of characters, informal way of writing, presence of slangs, emoticons and abbreviations are the important attributes of tweets.

The work by Go et al (2009) is the first research (as per their claim) carried out to determine polarity of tweets. They used machine learning approaches like Naive bayes, SVM and Maximum entropy classifier to determine the same. One of the basic requirements to do machine learning is the labelled data for training. In the case of large volume of tweets manual tagging is almost impossible. So they created training corpus by extracting tweets using Twitter search APIs and emoticons in the tweets to differentiate positive and negative tweets. Tweets with positive emoticons like ‘:)’ and negative emoticons like ‘:(‘ are separated to create a training corpus. The authors obtained good results with the three algorithms for the tweets from the domain like consumer products, companies and people. They have experimented with different features and made certain interesting observations. They observed that usage of POS-TAGS does not make any impact for the sentiment classification of tweets. They have got slight improvement in their results when they used both unigrams and bigrams as features. They observed that unigram model outperforms all other models when using SVM.
A similar kind of work by Pak & Paroubek (2010) also created Twitter corpus by collecting tweets using the method proposed by Go et al (2009). They also collected neutral tweets that corresponded to user accounts published by newspapers and magazines in North America. They performed statistical linguistic analysis to study the frequency distribution of words, POS-TAG features of the words. They build a sentiment classifier using the multinominal Naive Bayes classifier, SVM and CRF. In contrast to previous approach they observed that POS and n-grams will improve the performance of classifier.

Barbosa & Feng (2010) use two-point approach which first classifies messages as subjective and objective, and further classifies the subjective tweets as positive or negative. To create the training data they used existing Twitter sentiment sites like TweetFeel. They used various syntax features of tweets which includes retweet, hash tags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. Agarwal et al (2011) extended the approach by Barbosa & Feng (2010) by using real valued prior polarity and by combining prior polarity with POS. They classified tweets into positive, negative and neutral sentiment classes. They experimented with three models - a unigram model (baseline), a feature based model and a tree kernel based model which presented a new tree representation for tweets. By combining unigrams with their features and combining the features with the tree kernel they got better results. Davidov et al (2010) used 50 hash tags and 15 emoticons as sentiment labels in order to classify tweets into multiple calssess like #sarcasm, #damn, #humor, #angry etc. They utilized four basic feature types for sentiment classification: unigram features, ngram features, pattern features and punctuation features. They used K-nearest neighbour algorithm for the classification. They observed that for binary classification of smiley-labelled sentences is simple task compared to classification of hash tag labelled tweets.
Further the accuracy of classification can be substantially improved by considering feature types - punctuation, word and pattern features. They also observed that the addition of ngrams does not make any marginal increase in the accuracy.

In contrast to above approaches Kouloumpis et al (2011) used Twitter hash tags to create Twitter corpus. They selected the top hash tags that they felt would be most useful for identifying positive, negative and neutral tweets. They used different features (n-gram features, Lexicon features, Part-of-speech features, Micro-blogging features) for doing classification experiments. They observed that the best performance on the evaluation data comes from using the n-grams together with the lexicon features and the micro blogging features. In contrast to previous approach by Pak & Paroubek (2010) they observed that including the part of speech features will reduce the performance.

A tweet may contain sentiments on multiple targets. Jiang et al (2011) implemented target-dependent sentiment classification of tweets. To achieve this they extracted syntactic features (nouns, noun phrases etc.) referring to the target in a tweet. They used Pointwise Mutual Information (PMI) to identify the top K nouns and noun phrases which have the strongest association with the target. They made experiment with the tweets collected and annotated by their own (459 positive, 268 negative and 1,212 neutral tweets) about companies, products and celebrities.

One of the main problems with the machine learning approach in Twitter sentiment analysis is the availability of labelled data. The machine learning approaches may not give good results for a data set from a different domain. The lexicon based method depends on the sentiment size and nature of the sentiment lexicons. Zhang et al (2011) introduced a hybrid method which makes use of lexicon and machine learning method to analyse
sentiments. In this method they first applied lexicon based method to identify semantic orientation of tweets. Then they extracted additional opinionated indicators from the extracted results based on lexicon based methods. Additional opinionated tweets are identified using opinionated indicators. Training examples are created from the lexicon based method. In addition to that they applied special rules for handling comparative judgements, negation that can change the orientation of a phrase. In this way the number of labelled tweets is increased and later applied machine learning (SVM) for classification of tweets. But the accuracy of this method depends upon the opinionated tweets generated based on the lexicon based method.

Saif et al (2012) extracted certain semantic features of the entities to improve the accuracy of the classifier. For instance, the entities ‘iPhone’, ‘iPad’, ‘MacBook’ appeared more often in tweets of positive polarity and they are all mapped to the semantic concept PRODUCT/APPLE. They compared their results with the baseline unigram and POS features and found that the semantic feature model outperforms the other baseline model. The approach is suitable to analyse sentiments from the tweets related to consumer products, companies etc. and may not be able to extract semantic features from the general tweets.

Few works on TSA have proven the scalability in a distributed environment. A large scale distributed system for TSA was developed by Khuc et al (2012). Primary contribution of their approach is the construction of a opinion lexicon builder for Twitter sentiment analysis. Map-reduce framework along with HBase is used to implement the lexicon. They used lexicon based methods to find the sentiment score and later used the sentiment score itself as a feature to apply for machine learning approach. The lexicon size was small which consists only 2411 positive words/phrases, and 1018 negative words/phrases.
Chamlertwat et al (2012) analysed the consumer insight about various products from the tweets. They used lexicon based scoring method using SentiWordNet to find the polarity of tweets and later extracted features (brand, product details) to analyse sentiments on each brand and their features. They summarized the consumer insights about various smart phones.

Ghiassi et al (2013) created Twitter specific lexicon for brand related tweets. They applied feature reduction using ngram and statistical analysis to build the Twitter specific lexicon. Later they developed a dynamic artificial neural network model (DAN2) where the input is the vector of features. They observed that Twitter specific lexicon is more effective than the other variations. They further observed that the DAN2 model outperforms SVM variation. Their experiments are limited to Justin Beiber brand related Twitter corpus and require further study for tweets related to other brands.

Kontopoulos et al (2013) developed an algorithm for creation of domain ontology from tweets and later used this ontology to split each tweet into a set of aspects relevant to the product or topic. Then the sentiment analysis results in the assignment of a sentiment score to each distinct aspect. The methodology may be useful to find sentiments on brands or products since the creation of ontology for such particular brand is easy.

A recent work by Mostafa (2013) analysed consumer sentiments on more than sixteen global brands by analysing sentiment score of more than 3500 tweets. He used lexicon based method to find the sentiment score of each tweet. He used a predefined lexicon with 6800 adjectives (2006 positive words and 4783 negative words). He observed that 20% micro blogs mention the brand name so Twitter will be a useful source to analyse the sentiments on global brands.
Hamdan et al (2013) experimented with set of additional features to analyse Twitter sentiments. Unigram model is extended with features from DBpedia, WordNet and SentiWordNet. They observed that using lexicon dictionaries like SentiWordNet is more useful to enhance the performance.

Ranked WordNet graph method (RW-SWN) (Montejo-Raez et al 2014) is a recent domain independent and unsupervised solution to find the polarity of tweets. In this method they used SentiWordNet a domain independent lexicon to find the polarity of the words. In addition, they used a PageRank value for each synset of the term in the tweet. PageRank value is the weight assigned to each synset after performing random walk algorithm on WordNet synsets. Final sentiment score is the average of the product between the difference of positive and negative SentiWordNet scores, and the PageRank value obtained with the random walk algorithm. They observed that their unsupervised method is a good alternative to the supervised methods.

Few approaches have utilized user and link information associated with tweet to analyse sentiments. Using the Twitter follower graph might improve the polarity classification. Speriosu et al (2011) used label propagation with Twitter follower graph to improve the polarity classification. They compared their approach with standard lexicon-based methods and supervised classifier. They did not find overall gains from using the follower graph as implemented in their approach. Tan et al (2011) also used user information to predict the polarity of tweets. From the given topic and a user graph, where a relatively small proportion of the users have already been labelled, their task is to predict the labels of all the unlabelled users. In contrast to the method by Speriosu et al (2011) they demonstrated that user level sentiment analysis can be improved by incorporating link information like follower graph. They empirically confirmed that the probability of
sharing opinion between two users correlated with the connectedness in the social network. Hu et al (2013) proposed a sociological approach and utilized social relation among users to analyse Twitter sentiments. They build a mathematical model for message-message relation to integrate sentiment relations between messages in sentiment classification. Message-message relations are made by considering the similarity of the message sent by the same user, and the friendship of one user with other etc. They observed that considering social relations can improve sentiment classification of tweets. The approaches which make use of user information are possible if the social connection is available. That may not be possible especially in the case of public tweets.

2.4.2 Twitter Sentiment Analysis and User behaviour

Several studies empirically confirmed that Twitter sentiments and user behaviour are correlated. This section discusses certain important works carried out in this direction. O’Connor et al (2010) find out the correlation between public opinion derived from polls with sentiment measured from tweets during that period. They observed that tweets sentiments are leading indicators of polls. They calculated the sentiment score from the day-to-day tweets. A tweet is considered as positive if it contains any positive word and negative if it contains any negative word. They calculated score of each day, the ratio of positive versus negative messages on the topic and then they took the moving average over a window of the past $k$ days.

Bollen & Pepe (2011) demonstrated that social, political, cultural and economic events are correlated with public mood levels by analysing the tweets and the timeline of important political, cultural, social, economic, and natural events occurred between August 1 and December 20, 2008. They developed an extended version of a well established psychometric instrument, Profile of Mood States (POMS). POMS-scoring function maps each tweet to
six individual dimensions of mood, namely Tension, Depression, Anger, Vigour, Fatigue, and Confusion. They argued that sentiment analysis of tweets can be done efficiently by syntactic, term-based approach rather than using machine learning techniques and machine learning techniques will be effective only when large volume of training data is available.

Bollen et al (2011b) further investigated collective mood states derived from large-scale tweets and identified the correlation with the value of the Dow Jones Industrial Average (DJIA) over time. They used Opinion finder tool to measure the polarity of tweets and GPOMS (Google Profile of Mood States) to measure six different mood dimensions from text content. The resultant seven public mood time series extracted potentially different aspects of the public’s mood on a given day. Further they extracted a time series of daily DJIA closing-values from Yahoo! Finance. Later they established the hypothesis that public mood as measured GPOMS and Opinion Finder can predict the future DJIA values.

Abbasi et al (2012) have selected an online community which resembles a real world community in terms of race, language, religion etc. They extracted tweets related with Arab Spring to analyse the mood before and after the event. They collected around 35 million tweets related to Arab Spring event almost for all of the countries involved in the revolutions. Further they used popular blog posts and facebook posts to collect data. After statistical analysis they observed that Yemenis were more concerned about security whereas Egyptians were more concerned about revolution and freedom.

Choi et al (2014) analysed Twitter sentiments to track emotional state of users. They used the tool LIWC (Linguistic Inquiry Word Count) to calculate emotions in the text. They analysed the sentiments in hour basis and observed that users were more likely show positive emotions during mornings.
and evenings than during afternoons. Further they illustrated the way to use Twitter data to track the occurrence of breaking events. Trung & Jung (2014) built a fuzzy propagation model to demonstrate how information or tweets are propagated by considering existing friendship relationship network in the Twitter environment. They observed that tweets with emotional words are retweeted more than other tweets.

2.5 COMMUNITIES IN SOCIAL NETWORKS

When users or actors in a network interact more frequently with each other cohesive groups will be formed. These cohesive groups are called communities in social network. Identifying implicit communities (communities emerging from the interactions and activities of actors in the social media) has attracted increasing attention and has been used in various domain. Communities are not unique and they vary depending on the application of specific needs or condition to be satisfied by the community (Tang & Liu 2010a). Traditional community detection methods are roughly classified as node centric, group centric, network centric and hierarchy centric. These classifications are based on the criteria to be satisfied by the community.

In a node-centric community each node in the community should satisfy certain properties. It can be complete mutuality (Palla et al 2005), reachability of members (McClosky & Hicks 2009) etc.

Group-centric communities consider connections inside a group as whole. It can be the average density of the nodes in the group (Abello et al 2002). Network-centric approaches consider global topology of the network. The goal is to partition the network into disjoint sets. Several approaches are there in the literature which includes groups based on node similarity (Gibson et al 2005), Latent space model (Hoff, P D et al 2002, Handcock et al 2007),

Hierarchy-centric approaches build a hierarchical structure of communities based on network topology. There are mainly two types of hierarchical clustering: divisive (eg. Newman & Girvan 2004), and agglomerative (eg. (Blondel et al 2008)). Traditional community detection in social media is studied in detail by Tang & Liu (2010) and Papadopoulos et al (2011).

When an actor or object shares more than one community overlapping communities are formed. The problem of identifying same wavelength communities is derived from the results of sentiments on various trending topics. Solution to the problem finally boils down to the problem of identifying overlapping bicliques from the bipartite graph. This section reviews some of the important papers regarding overlapping communities and related to our area of research.

Palla et al (2005) developed Clique Percolation Method (CPM) is one of the important methods to find overlapping communities in networks. Their approach first finds all cliques of size \( k \) (user specified parameter) from the given network. Then create a clique graph where the edges in the graph are formed in such a way that an edge in the graph is present if the cliques share \( k-1 \) nodes. Each connected component in the clique graph will form a community. This is an efficient form of finding overlapping community from general graph but not suitable for large graph since the enumeration of all possible cliques for a particular value will be computationally intractable. The approach cannot be applied to a bipartite graph. Shen et al (2009) extended the approach by applying agglomerative clustering algorithm after finding out the maximal cliques. The approach is also applied to general undirected graphs.
Chen et al (2010) provided a game theoretic frame work to identify overlapping communities from social networks. They defined utility function on each agent for the set of communities user participates. They decompose the utility of the agent into two components. The first component is the gain of an agent joining the communities and the second component is the loss associated with the agent’s action of joining the communities. Agent’s utility is calculated as the difference between her gain function and loss function.

Certain recent works have utilized social network interactions and ratings provided by users to infer communities from the social network. Sachan et al (2012) proposed models that can discover communities based on the discussed topics, interaction types and the social connections among people. They experimented on existing community and in their model they assume that a user can belong to multiple communities and a community can discuss multiple topics. From this model they are able to discover both community interests and user interests based on the latent interactions and associations. Yang et al (2010) proposed to infer trust based social circles by combining available rating data and social network data. They observed that these trust circles can be used for recommendation. A recent work by Chuan et al (2013) proposed a genetic algorithm to identify overlapping communities from graph. They used genetic algorithm to detect link based communities and later converts link communities to node communities.

Social networks with two kinds of actors can be realized by bipartite graph. Identifying overlapping communities from bipartite graph is required for several kinds of applications. The methods varies depends of the application of specific needs. Madeira & Oliveira (2004) discussed several biclustering methods for the analysis of gene expression data obtained from microarray experiments. They mainly identified four types of biclusters for their requirement 1) Biclusters with constant values. 2) Biclusters with
constant values on rows or columns. 3) Biclusters with coherent values. 4) Biclusters with coherent evolutions. Du et al (2008) proposed a method BiTector which will identify overlapping biclusters. Their method will first enumerate all all maximal bicliques and from the maximal bicliques they later identified the overlapping nodes to form biclusters.

Wang et al (2010) provided a framework to identify overlapping biclusters of users and tags from online social media. They observed that the formed biclusters can determine explicitly who is interested in what and which is helpful in understanding and learning from the groups. Their method is able to find out user/tag group structure and their correspondence.

2.6 CONCLUSION

Identification of same wavelength communities is a problem derived from social science. Those who are having same thinking pattern have high probability to group together. Social scientists have identified several network parameters which induce homophily. Sentiments towards an issue or a topic are considered as one of the important dimensions which induce homophily. Twitter is considered as a booming online micro blogging service in which people express their sentiments on various issues. The solution to the problem mainly depends on the sentiment analysis of tweets. Tweets are basically ‘short texts’ and written usually in a cryptic and informal way. The review of literature discussed the major works on short text analysis and the approaches used to address the issues related with short texts in general. Our problem needs a domain independent solution for analysing sentiments of tweets since the tweets generated can be multifarious. We discussed the different approaches used for analysing sentiments and how the Twitter sentiments have been used for analysing user behaviour.
Same wavelength communities are groups formed on the basis of opinions and sentiments of similar hue towards various issues by different individuals. Communities are distinct and most often identification of communities depends on the application of specific needs. In our problem the same wavelength communities are perceived as bicliques and all bicliques from the bipartite graph needs to be identified. The review discussed the major approaches in the area of overlapping community detection in the social network and how the identification of such communities applied to different domain. The coming chapters discuss the methodology applied to solve the problem.