# Sentiment and Emotion Analysis for Context Sensitive Information Retrieval of Social Networking Sites: A Survey

D.I. George Amalarethinam, Ph.D
Director – MCA and Associate Professor of
Computer Science, Jamal Mohamed College,
Tiruchirappalli, Tamil Nadu, India.

V. Jude Nirmal
Assistant Professor, Department of IT,
St. Joseph's College (Autonomous),
Tiruchirappalli, Tamil Nadu, India

## ABSTRACT

Context Sensitive Information Retrieval (CSIR) is quite a challenging issue because of the complexities involved in the process from semantics and ontology to the huge amount of processing capacity required to make it possible in real time. Understanding the semantic gap (where context is neglected) plays a major role in elimination false positives and improving the true positives in the information retrieval process. With big data becoming ubiquitous due to the volume, velocity and variety of data being presented and analysed in almost all the domains today, context sensitive analysis and interpretation of big data becomes important.

This paper presents a comprehensive survey of the existing techniques for big data analysis based on massively parallel processing techniques like GPGPUs (CUDA), Hadoop Map-Reduce and also Data Warehousing. This paper presents a discussion about the datasets that are available for research and also the applications that could be thought of by context sensitive analysis of social media data. Also this paper provides research directions for context sensitive information retrieval and sentiment analysis in big data based on massively parallel processing architecture.

## Keywords

Context Sensitive Information Retrieval, Sentiment Analysis, Emotion Analysis, CUDA, Hadoop, Parallel mining

## 1. INTRODUCTION

The information age has brought about an explosion of data collected from various sources. These data are information rich, but it is difficult to efficiently process this data and extract valuable information from them. Information extraction was carried out on the basis of syntactic structures rather than semantics. The ability to extract information based on the context. Context Sensitive Information Retrieval has also gained prominence in the recent years. With the vast amount of data available, meaningful information extraction will help in a variety of ways which will be discussed in the later chapters.

Semantics is referred to as the study of meanings, in other words, it is meaningful computing. It uses Natural Language Processing to support the process of information retrieval. In order to extract meaningful information, a semantic system uses the content of search, location, word variation, intent of text, synonyms, concept matching and natural language queries. When providing a semantic based query the IR system analyzes the searcher's intent and the contextual meaning to provide more relevant results. The process of retrieving information from a document or a set of documents is usually performed by the process of querying. In context sensitive system a single query is not sufficient to retrieve the results, it may require more than one query. Hence a feedback mechanism is usually imposed on the system to ensure high accuracy levels. Shen et al in [1] proposed that Feedback incorporation in an information retrieval system can be either implicit or explicit. The explicit or relevance feedback requires the user to explicitly provide inputs regarding the process of retrieval. They will be intimated to rank documents, mark similar documents or categorize documents. Even though the above method [1] is effective, it is not often successful, because the users do not usually come forward to provide such information. The implicit feedback on the other hand evaluates the queries and the query modifications provided by the user. The user does not always get the required result in a single query. They perform many modifications in the queries to retrieve the desired result. These modifications can provide the essential information required for the system. In short, the implicit feedback exploits the available information by using the user's history data for its analysis.

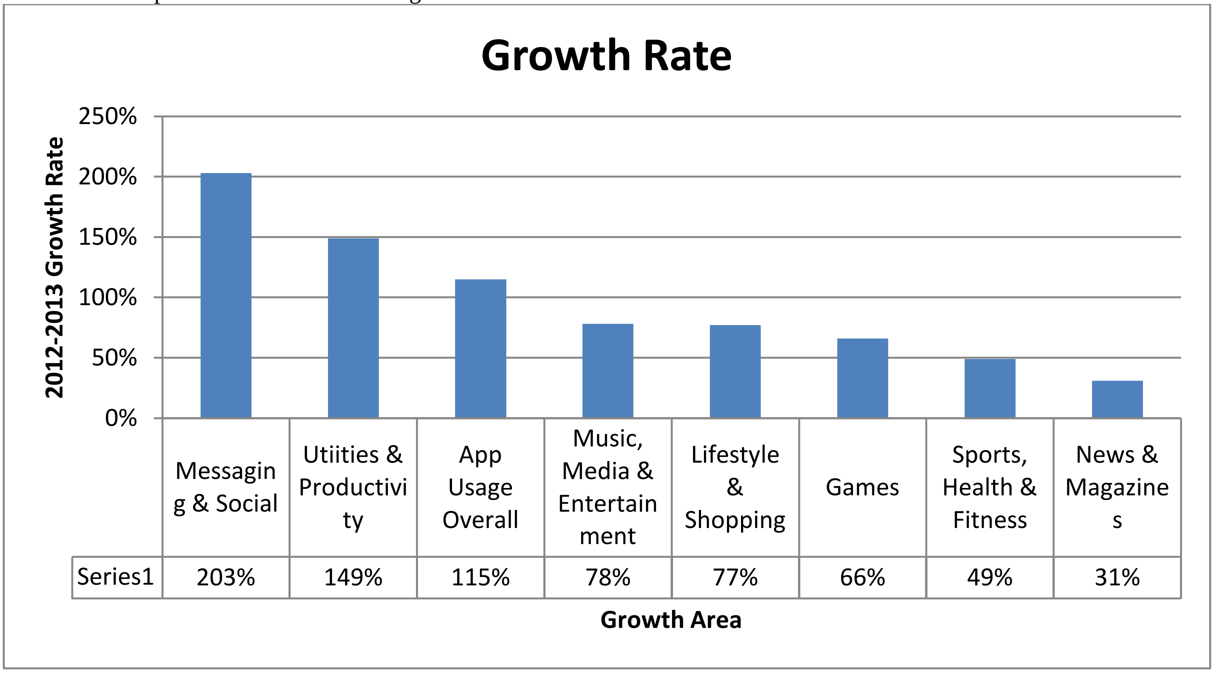
### 1.1. Multi-lingual Semantics

Language is the means of communication. Hence performing searches based on one's own language certainly has many advantages and will provide more accurate results. A language mediated search is not only tempting but also is effortless and effective. Web inherently supports multiple languages inherently, but there is no data integration mechanism present[2]. The problem is to bridge the gap between language specific information needs of the users and language independent semantic context, which can help in universal access of data. The most apparent hindrance is that ontologies are language specific. Hence incorporating multi-lingual searches at this point are difficult if not impossible.

### 1.2. Social Network Outburst

The first documented social networking site, launched in 1995 was classmates.com. As of November 2011, it boasts of 50,000,000 registered users. Fig 1 shows that the growth of social networking has been increased about 203% in 2012-2013. There are many undocumented sites, hence the beginning of social networking cannot be accurately dated. Some of the social networking sites that gained prominence are Friendster, Hi5 and LinkedIn, that were launched during (2002-2003). The actual outburst occurs after the introduction of MySpace, Orkut and Facebook (2003-2004). And now

Google+ (launched in 2011) boasts of 500,000,000 users, which shows the impact that social networking sites have on the masses.



| | Messaging & Social | Utiities & Productivity | App Usage Overall | Music, Media & Entertainment | Lifestyle & Shopping | Games | Sports, Health & Fitness | News & Magazines |
|---|---|---|---|---|---|---|---|---|
| Series1 | 203% | 149% | 115% | 78% | 77% | 66% | 49% | 31% |

**Fig 1: Growth Rate of Various Services**

Further, most of the users are willing to share a level of their private information (public profile of users) to the public. This information proves to be very significant in most decision making and analysis processes [31].

### *1.2.1. Social Networking : An Overview*

#### *1.2.1.1. Facebook*

Facebook, one of the top social networking sites was launched in February 2004. The users create profiles describing themselves. These profiles can be linked to the profiles of their friends. Any friend request is validated by the request receiver. Hence in Facebook, all friendships are reciprocated ties. The profiles are stored as nodes and it contains undirected edges to all the nodes(profiles) that are marked as friends. The entire Facebook can be considered as an undirected graph, while 'like' and 'follow' options create directed edges. It used both text and image formats for communication. Videos can be uploaded for viewers. Fig 2 shows the number of monthly users of Facebook from 2004 to 2013.

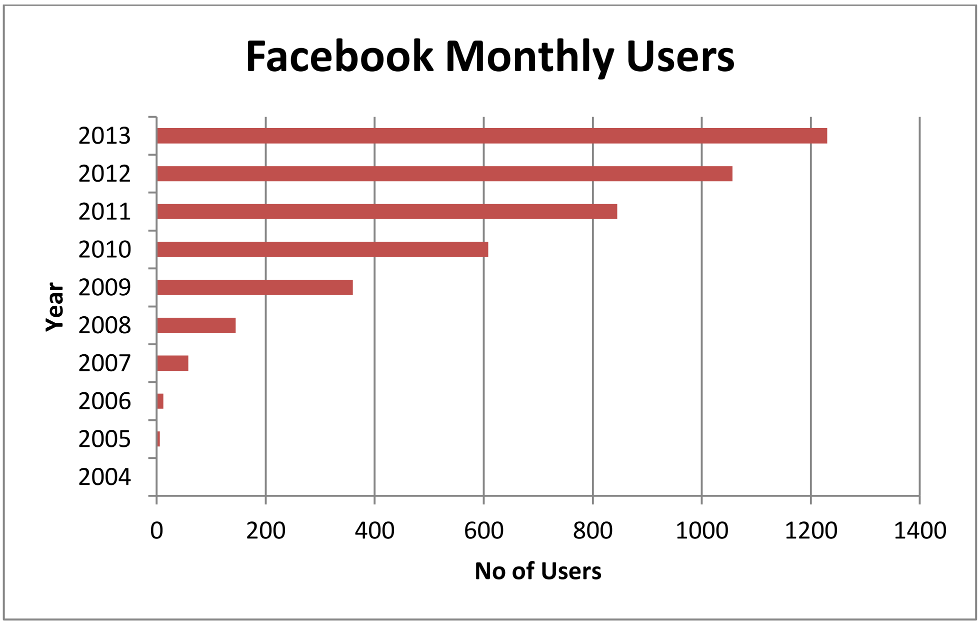


**Fig 2: Facebook Monthly Users**

#### *1.2.1.2. Google+*

Google+ was launched in June 2011. It also manages the profiles of the user. The advantage of Google+ is that the user profile is integrated into other Google services. It includes basic personal profiles, identity service selections and helps in organizing people into groups or lists for sharing. News categorizations are available for users to follow their interests. It provides a mechanism to 'follow' people or add them to your circles. News sharing is available along with an option to video chat (hangouts) with 10 friends (max). It uses a directed graph to represent relations.

#### *1.2.1.3. LinkedIn*

This business oriented networking site was launched in December 2002. It helps to create professional networks, hence a person's academic and job profiles play a major role here. It serves as an intermediary between the employee and the employer. It helps maintain contact details, and supports 20 languages. Invitees are also intimated about their followers and they have the advantage of setting the inviter's status. A person can either be added as a connection or just a follower. It also maintains a directed graph for managing the profiles.

#### *1.2.1.4. Pinterest*

This visual discovery tool was created to collect ideas and follow people with similar interests. It was launched on March 2010. It is mostly a visual bookmarking of web pages and images. Not much of a user's profile is managed here, instead a person is identified by the boards(interests/themes) that they manage. A board is considered to be a place where various related bookmarks are placed and managed by the user. A user can create several boards and there are provisions to create hidden boards. Hence data management here is mostly organization and categorization of url links. An interest graph is created between the two boards rather than to users.

#### *1.2.1.5. Twitter*

Twitter, the micro-blogging site was launched on July 2006. Users can send small messages restricted to 140 characters called tweets. A registered user can both read and post tweets, while an unregistered user will only be able to read tweets. People can follow their interests, their friends or famous personalities. Provision to block a user from following is also available. Hence a twitter graph is directed graph with one to many connections. It uses only textual data and sometimes URL's.

## 2. SENTIMENT AND EMOTION ANALYSIS

Sentiment analysis is the process of extracting the contextual polarity from a text. A document polarity can be positive, negative or neutral. The process of sentiment analysis is performed by initially identifying the adjectives and adverbs in the text. The semantic orientation of the identified word is analyzed and a class is assigned to it accordingly. The average semantic orientation is taken as the classification for the entire text. The analysis is entirely performed on the basis of the single identified words. This analysis can be performed both manually and can also be automated. An extension of this system can be thought of as a scaling system, that grades the level of polarity. In general, these scales range from -10 to 10. But this alone is not sufficient. The qualitative description of the emotion is mandatory to determine the appropriate context. The next level of sentiment analysis is the Emotion Analysis, that describes the sentiment beyond the level of polarity. It provides an emotional description of the sentiment level.

The importance of sentiment analysis arises due to the increase in the usage of social networking medium and information sharing (as discussed in Section 2.2). User's interactions in the social media are in the form of likes or comments or follows. These depict the interest of the user. These data are public and the users do not take trouble in making the private. Hence mining can be performed using this open data. This opportunity has given rise to an increase in the online business. The concentration of the business is always on the content and the context of the text, rather than just the content, which can be deceptive.

## 3. RELATED WORKS

### 3.1. Sentiment Analysis

A semantic based review classification technique is discussed by Turney et al [3]. It uses the methodology of Unsupervised learning. The PMI-IR (Pointwise Mutual Information-Information Retrieval) algorithm is used for processing. It selects two consecutive words as opposed to single words selected by most of the methods. The semantic orientation of these words are then calculated using the PMI-IR technique. Finally, the average semantic orientation of the entire text is calculated and the text is classified as recommended or not recommended. It uses the Alta Vista engine for its processing, hence the major downside of this approach is high time consumption.

An analysis of the machine learning methods in the process of sentiment mining is performed by Pang et al [4]. It uses Naive Bayes, Maximum Entropy Classification and Support Vector Machines for performing the text classification. The IMDb comments are used as the base data for processing. The advantage of this method is that it does not use stemming or stoplist for processing. This paper concludes that SVM works best in the IMDb data, while the others closely follow.

Fuzzy semantic based emotion mining is presented by Loia et al [5]. It uses fuzzy entities for describing the parameters. But membership functions for the fuzzy entities are not defined. It performs Opinion Mining along with Sentiment Analysis. Due to the usage of AI based mining, parallelization could prove to be very efficient. Triangular fuzzy membership is almost similar to the six point rating. -3, -2, -1, 0, 1, 2, 3. Though fuzzy systems are considered to be continuous , this doesn't help much in the real sense. Ghazi et al [6] presents an algorithm to extract emotions from a sentential context. But this method ignores second order logic and emotions. Usually emotional elements are expressed in complex forms and not simple sentences. It uses SVM and Logistic regression to perform the classification process. Li et al [7] considers using artificial intelligence or Machine learning approaches for the process of sentiment analysis. Deep Learning systems could uncover new hidden patterns. This method could be utilized by the retail industry as well as by social networking sites for Ads and targeted marketing.

The effect of ensemble methods for sentiment classification is discussed by Wang et al [8]. Ensemble refers to a combination of methods rather than a single method. This method contains multiple learners for the same problem. It proposes a set of hypothesis instead of a single hypothesis. The initial process deals with extraction of features and the next step converts these features into feature vectors. The feature representation method used here is the Bag of Words framework. Classifier training is performed on the vectors using various classification approaches. Testing is performed using the unseen data. This method leads to high computational load, hence parallel implementation would be a better option.

A method to test the validity of the sentiment classification is proposed by Wijnhoven et al [9]. When analyzing a data, problems can be created due to bias. This can be due to the following reasons; mismatch in the demographics of the reviewers, bias due to previous experience or manipulation of reviews. These are classified into demographic, event or manipulation bias. Tests for each of the bias types are discussed and the bias elimination methods are proposed. It helps in identifying the usefulness of Sentiment Analysis.

A social emotion detection for online news is described by Lei et al [10]. It performs the social emotion detection of online users in the online news domain. Social emotion refers to user's response when exposed to news articles. It performs the process in three phases. Document selection is initially performed to choose a well formed training set. The part of speech information is used for tagging and feature set extraction, finally it generates the social emotion lexicon to predict the emotions.

A method used for emotion analysis of content in mails and books is proposed by Mohammad et al [11]. It creates an emotion lexicon and text analysis is performed on the basis of this lexicon. It deals with gender based emotions. Various texts are considered and the emotion level in them are analyzed and depicted graphically.

## 3.2. Social Networking based Sentiment Analysis

The effect of social networking using LinkedIn, the business based networking site is discussed by Bradbury et al [12]. It describes a scenario, and using which a variety of ways in which a person's information can be gathered. The probability of security threats, finding an individual's area of work in the corporate scenario, finding influential individuals or Top- k-influential nodes are only a very few activities that can be performed on the data.

A ranked sentiment polarity classification algorithm is presented by Raez et al [13]. Twitter data in the form of ranked WordNet graph is used for the classification process. A vector of weighted nodes is extracted from the WordNet. Polarity score is provided by the SentiWordNet. The Random Walk algorithm weights the sunset from the text with the polarity score. This method has proven to be unsupervised and domain independent and offers performance comparable to SVM.

A content based behaviour analysis is presented by Lee et al [14], as opposed to the traditional structural analysis of social networks. This method helps identify user roles using their behaviour and hence deduces role change patterns. The advantage of this approach is that, it deduces the user's attitude and interest in the process. The downside of this approach is that it does not consider the semantics of the data for performing the analysis. It can help us model group or consensus behaviour based on roles, with or without leadership groups. Structural analysis could help us classify people into predefined roles. Dynamic nature of web 2.0 could not be mapped using structural elements alone.

A framework for opinion mining in Twitter is proposed by Khan et al [15]. Sentiment Analysis in this domain can be performed to understand public's feelings towards brand, business, directors etc. The general problem encountered during the process of sentiment analysis is the accuracy of the classification, dealing with data sparsity, sarcasm etc. Khan et al [15] proposes a hybrid approach to classification. The method is carried out in three phases. The initial pre-processing phase performs the analysis for slangs and abbreviations. This method is then preceeded by the actual classification technique. The emoticon analysis and SentiWordNet analysis are performed along with detection of positive and negative words. The final evaluation phase analyzes the results in terms of Accuracy, Precision, Recall and F-Measure. Each of the preprocessing steps is compute intensive and real time analysis of a stream is highly intensive. Parallel Architectures could make this usable. Machine Learning system would perform better.

A survey of mining techniques that can be used on the social networking data, based on the graph structure are presented by Nettleton et al [16]. It provides a list of datasets that can be used for studying online social networks. It also provides the metrics and measurements for the analysis. List of graph based Data Mining algorithms are discussed here. Lists out and explains pre-processing and sampling of graphs. Provides an overview of the influence and recommender systems. Algorithms for clustering or community detection in OSN are discussed here. Behaviour Analysis is performed and the Information diffusion process in OSN is described in detail. It also discusses the relationships and how it affects behaviours in OSNs.

The process of predicting a user's personality by using the persons interactions in Facebook is described by Ortigosa et al [17]. These prediction mechanisms can be used in various areas such as adaptive applications. The Facebook dataset is taken and is applied on different Classification methods and the results are cross referenced and merged to obtain the tendency of a person. This method is best suited for identifying abnormal interactions. Terrorist activities could be isolated. Novelty detection could be done to isolated unique personalities. The C 4.5/J48 algorithms available in WEKA are used for Classification.

A case study of Twitter is presented by Jung et al [19]. For security analysis, we can combine micro-texts from several social networks. The resultant value could also help in better marketing and brand analysis. The resultant data will help determine the user's profile and their interests. It helps determine patterns, hence in turn can be useful in many ways such as determining terrorist activity. Information sharing can be performed using these microtexts. Context based mining can help discover previously unknown patterns. When considering the field of marketing, these texts can help determine brands and products of interest. Determination of opinions about politicians can be performed. We can also determine data about events, about news, about organizations and celebrities.

A preference based mining technique is discussed by Zhou et al [20]. This method helps determine major contributing nodes in a network. Quality alone is considered as the parameter for determining the rank of a node. Title alone for user preference consideration. This might prove to be a downside, because titles could be stylized and Biased. Negative impact is not considered. The current method uses a greedy method to arrive at the solution. Uniform distribution is just altered with weights. Monte Carlo -simulation is used in the process. Fuzzy nature of the influence of several nodes on a single node is not considered. KamelBoulos et al [22] discusses the application perspective of mining social data in health care, environmental and national security domains. Still not so mature ,but future research directions are promising. Jang et al [23] discusses an information fusion approach for analysis of business ads."Deep" is used in the context of fuzzy or degree. Not really a Deep Learning system. Applicability for real world scenarios must be discussed further. An agent based social network community mining is discussed by Huang et al [24].Generalizes the mining to a distributed constraint Satisfaction problem. The mechanism is decentralized and self-organizing or aggregating. This method is highly concurrent and asynchronous, so could be effectively parallelized on GPGPUs.

## 3.3. Parallelization in Sentiment / Emotion analysis

### *3.3.1. GPGPUs for Parallelization*

A survey of Parallelization of Machine Learning approaches is described by Sujatha [18]. It discusses the Machine learning techniques, their advantages and disadvantages and Parallel implementations of these Algorithms. In general, parallel based implementation can be GPGPU based or using Map-Reduce. Both the mechanisms are discussed here.

Mechanisms for parallel processing of large graphs is presented by Kajdanowicz et al [21]. This method provides methodologies for processing graphs in a parallel environment.GPU implementation of the problem could be faster assuming enough GPGPUs with a capacity to hold the graphs in memory. BSP is better than Map-Reduce for systems that fits into local memory. Map-Reduce can still be utilized for enormous systems with very large data structure requirement.

Lobachev et al [26] provides a mechanism to estimate performance of massively parallel code. It also allows exact measurement and prediction with Scalability in terms of problem size and processor members. This paper provides a new metric to understand parallel program quality. Applicable to multi-core systems and peta scale super computers as well .

### *3.3.2. Map-Reduce for Parallelization*

A parallel and incremental classifier is discussed by Qing et al [25]. SVM is used for classifying highly imbalanced and Textual data. This proves to be very effective for social network data analysis. This is based on Hadoop Map-Reduce. This method is limited by the Batch-processing nature of MapReduce. No interactive mechanisms. A GPU based implementations can still prove to be a better option.

A computational model for MapReduce is presented by Karloff et al [29]. It concentrates on reducing the total memory per machine and the total number of machines. These are placed as the constraints in the algorithm, hence the system uses minimum resources. The time constraint is not set to linear, hence the algorithm runs in polynomial time.

A HPC compatible MapReduce Framework is proposed by Fadika et al [30]. MapReduce was designed to work in Hadoop File System(HDFS). It is in general, not compatible with most of the existing HPC environments. Fadika et al [30] presents a MapReduce implementation that is suitable for most of the HPC environments. It also discussed the design choices for developing an interoperable framework. This method exhibits better performance gains.

# 4. APPLICATIONS OF CONTEXT SENSITIVE INFORMATION RETRIEVAL IN SOCIAL MEDIA

There are several applications of context sensitive information retrieval in the social media. Several other possibilities are being explored and commercialized now. With a huge mass of people involved in those activities, almost there is no limit for the amount of personal and private information that is being shared everyday without our knowledge. From job profile matching by Human Resource Recruiters to Home Land Security, everybody is analysing the social media information for understanding and identifying the required patterns.

## 4.1. Job Profile Matching, Job Search

One of the latest and emerging field and uses of the social media has been Human resource recruitment. Offshore accounts are monitored and their patterns are analyzed and service calls are made to those selective individuals requesting them whether they are interested in switching over to another company in the same country or other. If they are interested they process the request further.

This process is further strengthened by the current social media context as everything including their profile is being shared and also the context sensitive analysis could even suggest their current loyalty level and state of attachment towards the current employer. With such information, the act of pursuing them to switch companies will be a walk in the park. On the other hand the companies can do this kind of analysis in order to identify the potential employees who are likely to leave and then encourage them with incentives and even counselling if necessary.

### *4.1.1. Problems in Data Handling*

The problem with this type of analysis is that the content is going to be in a wide variety of formats and also in various languages. Multilingual context sensitive analysis is one of the research issues and further the various formats in which a resume or profile will be must be taken into consideration. Even the profile information of two persons applying for the same job need not be the same. There could be a different set of documents related to each of their profile.

We need column store technology and also big data processing environment in order to accomplish it. Document databases would be the choice because of its flexibility in storing a wide variety of data. Map-Reduce based processing of information would be the most suitable form of analysis as it is massively parallel and also well suited for this semi-structured data handling.

## 4.2. Security: threat to national and home land security

Content filtering can be used in real time to avoid communal and regional clashes due to inappropriate content sharing in social networks. They can be used to Monitor Terrorist Activity and Content Sharing in the Social Media. Pattern matching techniques can be used to identify illegal content sharing in the social networks. Searching for such patterns will lead to identification of illegal content. The analysis of what influences people to make decisions can be performed based on their context. Context sensitive analysis of the scenario can be used to understand the decision making process. Further the complexity in analysis is due to the multilingual and multimedia content available in today's content sharing networks.

### *4.2.1. Problems in Data Handling*

Real time service constraints will pose a major problem in handling the data. The system needs to filter the content before it is delivered to others. This gives a processing time of not more than a few seconds. Need to analyze the nature of content in that context and make decisions which involve semantics and machine learning. It will be a time consuming process and hence parallelization options to be considered. It is not possible for a single authority to delete or block the contents posted by people from various geographic locations as they are bound by different laws. One article which could be a threat to the integrity of India will not be of much importance in US and vice versa. Collaborative operations will involve communication and hence bottlenecks.

## 4.3. Effective content sharing : Facebook and Twitter, page suggestions, friend suggestions

Privacy protection should be imposed in accessing the data from social networks. Location based services provide a major turn down. Android applications in mobile could expose locations unintentionally. Other privacy issues arise in sharing the content of the messages and sharing the friends list and the groups in which one is having membership etc. Services used by a particular user must be kept confidential. Combining location based services and web content mining could lead to privacy issues.

Friend suggestions based on location, place of study, place of work etc are being done today. This model could be extended effectively to areas of interest and emotional values based friend suggestion system if context sensitive analysis is possible. Identifying that part of the graph that has more connections in terms of its semantics and patterns of interest could lead to such a model for effective friend suggestions and content sharing. Though it is viable, privacy and security issues have to be addressed in such a scenario.

Clickstream data analysis especially from the mobile phones could lead to privacy issues. All these are context sensitive in that if a click stream analysis and information fusion is performed on a website accessed through the PC or Laptop then there is no problem of exposing the location but the same, when done in the context of mobile access then it could lead to privacy concerns.

Peer pressure is one another factor concerning the social media applications. Targeted marketing becomes easy as the peer pressure increases as people share everything they have and own and whatever they feel is best. Now from friends and family we are being influenced by people from different geographical locations and from varying cultures which pose a threat to the right or freedom of thought or decision making. All our decisions are now being motivated by the social media applications and the online society as such.

### *4.3.1. Problems in Data Handling*

Major issues in handling the data from various content sharing platforms is that the data representations do vary. Face book allows us to post content that could be anyway from a single word to a huge paragraph, and also allows us to add video and images at almost all possible places. Twitter on the other hand allows us to send tweets which are inherently limited to 140 characters. Linked-in in a kind of network for professionals which is more towards their public rather than personal information. Also the nature of content in the network does differ and so are the connections in between the users. There are undirected graphs like facebook, directed ones like twitter etc. The way they must be stored differs to a greater extent and at the same time the volume and velocity or rate of information explosion is high. So it mandates a NoSQL based big data environment for storage and effective retrieval of this information. But performing context sensitive information filtering and analysis on this kind of data is a research challenge.

## 4.4. Targeted Marketing/Advertising

The amount of personal preferences that we share every day is enormous. Actually the information that we share through like or dislike in a social networking environment or whom we follow in twitter environment or where do we work and such information could be effectively used for targeted marketing. Even Ad companies use this social followers' count to determine the popularity of individuals before signing up with them for ads.

The popularity of celebrities is nowadays determined by their followers count. This information could be mined for effective targeting of individuals based on their area of work, and type of equipment that they may require and like and the range in which they could afford to buy etc,. Such information is used to provide specialized and targeted advertisement links in the social media itself. Though the information is anonymous it is very rich in terms of the information that they can obtain. With context sensitive information retrieval and by combining the information from various content sharing sites, targeted marketing can go a long way.

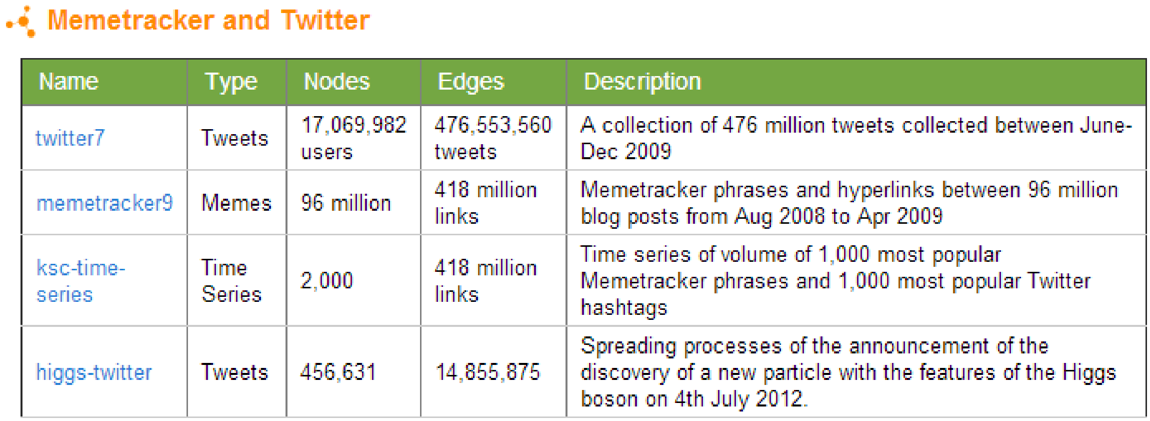
The information obtained from these sites to be used for targeted marketing must be analysed properly before using it effectively, as the information is about a wide variety of people from varying geographical regions and from various cultures. Culture and location specific marketing models must be constructed before we can effectively use this information for targeting the individuals.

# 5. STANFORD LARGE NETWORK DATASET COLLECTION (SNAP)

SNAP was developed in 2004 and is largely growing and contributes in the area of analysis of large social and information networks. It contains a large collection of standard datasets for use in the social networking domain.

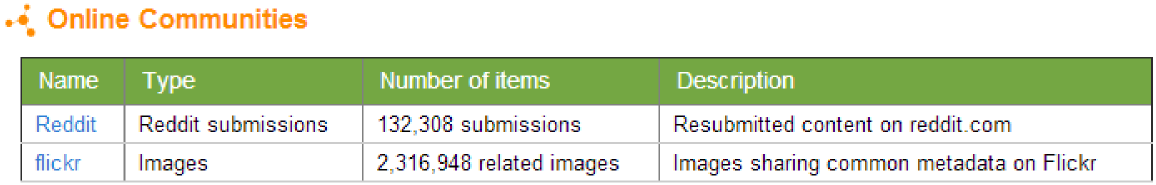
The areas include [27]:

- Social networks: online social networks data, where nodes represent people's profiles and edges represent interactions between people. These sections contain data from Facebook, Gplus, Twitter, Epinions, Live Journal, Pokec, SlashDot and wiki vote.



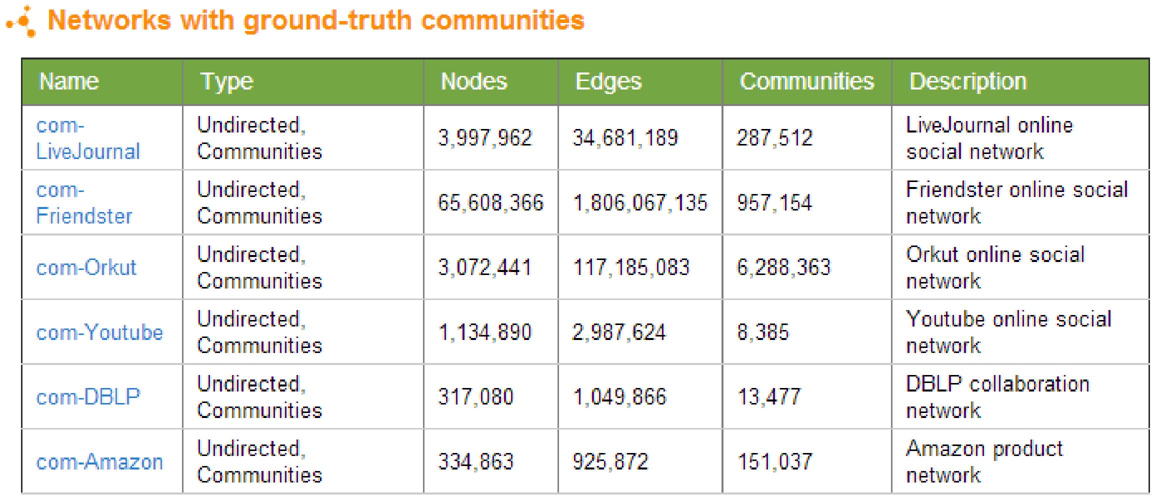
### Memetracker and Twitter

| Name | Type | Nodes | Edges | Description |
|---|---|---|---|---|
| twitter7 | Tweets | 17,069,982 users | 476,553,560 tweets | A collection of 476 million tweets collected between June-Dec 2009 |
| memetracker9 | Memes | 96 million | 418 million links | Memetracker phrases and hyperlinks between 96 million blog posts from Aug 2008 to Apr 2009 |
| ksc-time-series | Time Series | 2,000 | 418 million links | Time series of volume of 1,000 most popular Memetracker phrases and 1,000 most popular Twitter hashtags |
| higgs-twitter | Tweets | 456,631 | 14,855,875 | Spreading processes of the announcement of the discovery of a new particle with the features of the Higgs boson on 4th July 2012. |



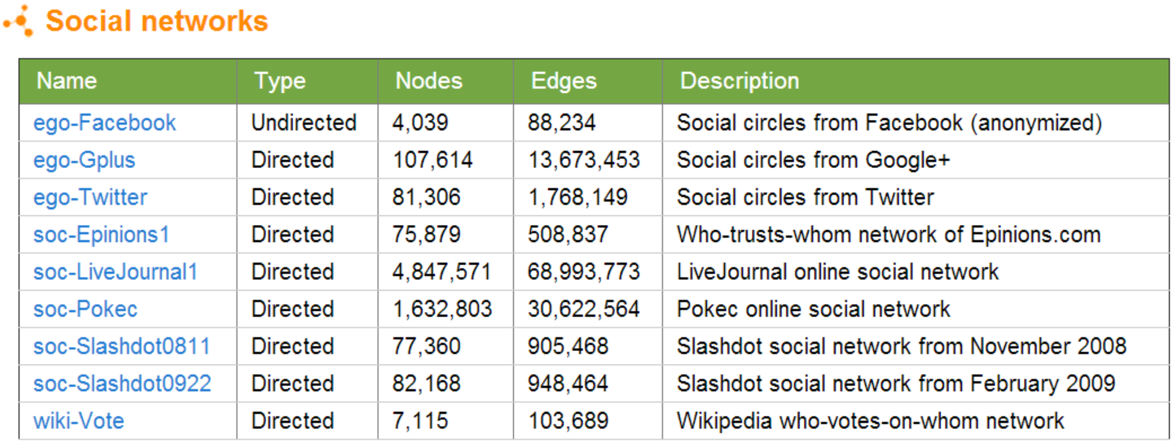
### Online Communities

| Name | Type | Number of items | Description |
|---|---|---|---|
| Reddit | Reddit submissions | 132,308 submissions | Resubmitted content on reddit.com |
| flickr | Images | 2,316,948 related images | Images sharing common metadata on Flickr |



### Networks with ground-truth communities

| Name | Type | Nodes | Edges | Communities | Description |
|---|---|---|---|---|---|
| com-LiveJournal | Undirected, Communities | 3,997,962 | 34,681,189 | 287,512 | LiveJournal online social network |
| com-Friendster | Undirected, Communities | 65,608,366 | 1,806,067,135 | 957,154 | Friendster online social network |
| com-Orkut | Undirected, Communities | 3,072,441 | 117,185,083 | 6,288,363 | Orkut online social network |
| com-Youtube | Undirected, Communities | 1,134,890 | 2,987,624 | 8,385 | Youtube online social network |
| com-DBLP | Undirected, Communities | 317,080 | 1,049,866 | 13,477 | DBLP collaboration network |
| com-Amazon | Undirected, Communities | 334,863 | 925,872 | 151,037 | Amazon product network |



### Social networks

| Name | Type | Nodes | Edges | Description |
|---|---|---|---|---|
| ego-Facebook | Undirected | 4,039 | 88,234 | Social circles from Facebook (anonymized) |
| ego-Gplus | Directed | 107,614 | 13,673,453 | Social circles from Google+ |
| ego-Twitter | Directed | 81,306 | 1,768,149 | Social circles from Twitter |
| soc-Epinions1 | Directed | 75,879 | 508,837 | Who-trusts-whom network of Epinions.com |
| soc-LiveJournal1 | Directed | 4,847,571 | 68,993,773 | LiveJournal online social network |
| soc-Pokec | Directed | 1,632,803 | 30,622,564 | Pokec online social network |
| soc-Slashdot0811 | Directed | 77,360 | 905,468 | Slashdot social network from November 2008 |
| soc-Slashdot0922 | Directed | 82,168 | 948,464 | Slashdot social network from February 2009 |
| wiki-Vote | Directed | 7,115 | 103,689 | Wikipedia who-votes-on-whom network |

- Networks with ground-truth communities :these correspond to data similar to orkut, facebook, myspace etc. They contain data from LiveJournal, Friendster, Orkut, YouTube, DBLP and Amazon.
- Citation networks : nodes represent papers, edges represent citations. It contains data from HepPh, HepTh and Patents.
- Communication networks : email communication networks whose edges correspond to the communication carried out between two individuals. It contains data from EuAll, Enron and Talk.
- Collaboration networks :this refers to co-authoring details, where nodes represent scientists, edges represent other scientists who

are co-authors for the paper. It contains data from AstroPh, HepPh, HepTh etc.

- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Web graphs : nodes represent webpages and edges are hyperlinks. They constitute graphs from Google, Stanford, NotreDame and Berkley domains.
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Internet networks : nodes represent computers and edges communication. A sequence of snapshots of the Gnutella peer-to-peer file sharing network is shared here.
- Road networks : nodes represent intersections and edges roads connecting the intersections. Road networks of California, Pennsylvania and Texas are shared here.
- Autonomous systems :The graph of routers comprising the Internet can be organized into sub-graphs called Autonomous Systems (AS). Each AS exchanges traffic flows with some neighbours (peers). A communication network is created using those logs.
- Online communities : Data from online communities such as Reddit and Flickr
- Signed networks : networks with positive and negative edges, which represent positive or negative emotions. These are trust based connection in the social networks
- Location-based online social networks : Social networks dealing with geographical information. Data constitute from Gowalla and Brightkite.
- Wikipedia networks and metadata : Article submission, editing and voting data from Wikipedia
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

Data from this method is also available to the user in the form of UI Sparse matrix, or as Visualizations. The network types supported are Directed, Undirected, Bipartite, Multigraph, Temporal and Labelled. It also provides many statistics about the datasets such as number of nodes, number of edges, number of nodes in the largest weakly connected and strongly connected component, number of edges in the largest weakly connected and strongly connected component, average clustering coefficient, number of triples of connected nodes, number of connected triples of nodes, maximum undirected shortest path length and 90-th percentile of undirected shortest path length distribution.

## 6. MPPENVIRONMENTS CUDA, HADOOP, MAPREDUCE, DATAWAREHOUSING : A DISCUSSION

### 6.1. CUDA

CUDA™ [28] is a parallel computing platform and programming model invented by NVIDIA. It enables dramatic increases in computing performance by harnessing the power of the Graphics Processing Unit (GPU).Using CUDA, the GPUs can be used for general purpose processing and not exclusively for graphics rendering. This approach is known as GPGPU. The main advantage of GPU's is that even though they are not exceptionally fast in processing single threads, they have the capability to process multiple threads at the same time. Hence the amount of work completed is higher when compared to CPU's. GPU's in general can process large blocks of data efficiently. They support scattered reads i.e. code can read from arbitrary addresses in memory. CUDA 6 supports Unified Virtual Memory and Shared memory, which can be shared amongst threads.

### 6.2. Apache Hadoop

Apache Hadoop is an open-source software framework for storing and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. It is an open source implementation of Google's MapReduce and Google FileSystem (GFS). Allows for scale-out processing of petabyte scale data. It also provides distributed storage. It can work against flat files or certain database formats. Native processing involves imperative Java code. Other languages are supported through Streaming.

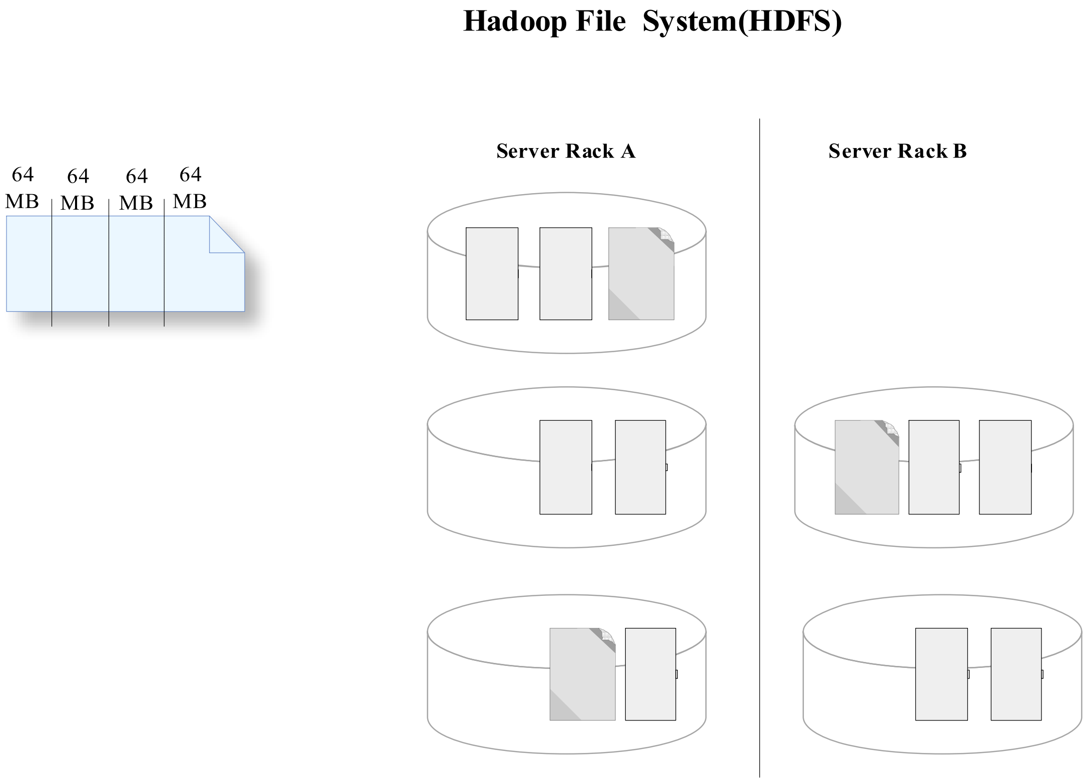


**Fig 3: Hadoop File System (HDFS)**

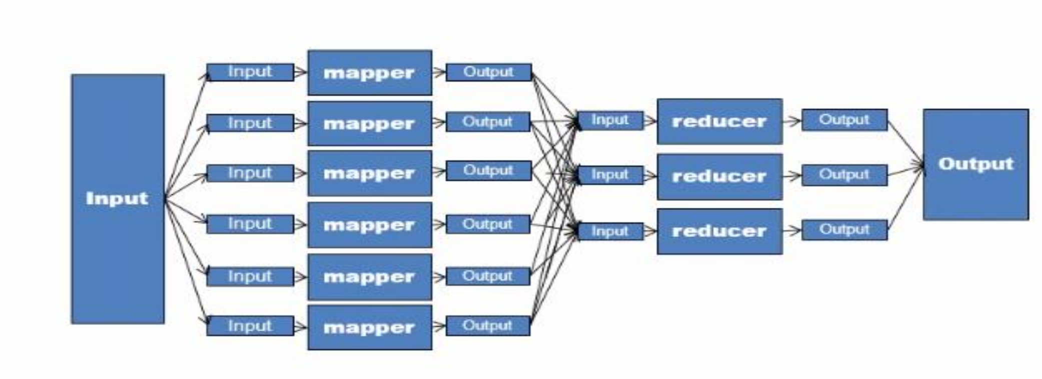


**Fig 4: Map Reduce**

Fundamental assumption made by Hadoop File System (HDFS) is that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework. Based on **Google Map Reduce** and **Google File System** Whitepapers. HDFS is a distributed, scalable, fault tolerant and portable file-system written in Java for the Hadoop framework.

## Map Reduce

A mapper is assigned to each block of data. Input and output in each phase are presented in the form of Key-value pairs. The Keys must implement WritableComparable interface. Shuffle and Sort plays a major role in solving the problem.

**Hive** provides an SQL like abstraction over Map Reduce. It opens up Big Data to masses. Development using MapReduce is time consuming. Further, it requires intimate knowledge of the framework. Hive was introduced to overcome these complexities.

## 6.3. Data Warehousing

Traditionally, a relational database comprises of a dimensional schema. This is optimized for reading data and not for writing data. An Enterprise Data Warehouse(EDW) acts as a central repository for the entire organization. An EDW is a database used for reporting and data analysis. A data warehouse creates a central repository of data by integrating data from various sources.

## 7. RESEARCH ISSUES TO BE ADDRESSED IN CONTEXT SENSITIVE ANALYSIS OF BIG DATA

## 7.1. Big Data

Big data is the term for a collection of large data set and complex structures. Due to its hugeness, it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges in processing big data includes capturing, curation, storage, efficient search, effective sharing, transfer, analysis, and visualization. Big data is of size 100s of TB to PBor higher. It involves data from finance, sensors, web logs, social media etc. A data is considered Big if the processing of datasets is too large for transactional databases, which involves Volume, Velocity and Variety. It is often advantageous to involve parallel processing in these areas.

## 7.2. Mining Big Data

As of 2012, the maximum size of data that can be processed in a reasonable amount of time was in the order of exabytes of data. The limitations also affect other areas like internet searches, business etc.. Data sets grow in size as they are increasingly being gathered by various input agents such as mobile devices, software logs, microphones ,cameras and wireless sensor networks. As of 2012, every day 2.5 exabytes of data were created.

Big data cannot be easily processed with most existing relational database management systems and desktop statistics and visualization packages, instead they require "massively parallel software running on tens, hundreds, or even thousands of servers“. Big data usually includes data whose sizes extend beyond the ability of commonly used software tools for manipulation. Big data sizes are a constantly altering themselves, as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set.

Mining such datasets which are unstructured and huge provides a lot of potential for the massively parallel processing environments like CUDA, Map-Reduce etc,. Several new apache projects for parallel mining algorithms targeting big data are in their incubation state. The need for such real time query and visualization of big data is much important now than ever before.

## 7.3. Information fusion and privacy issues in big data environment

In a collection of heterogeneous and autonomous information sources, we need to provide a system that allows its users to perceive the entire collection as a single source, query it transparently, and receive a single, unambiguous answer. Heterogeneous information sources are sources with possibly different data models, schemas, data representations, and interfaces. Autonomous information sources are sources that were developed independently of each other, and are maintained by different organizations, that may wish to retain control over their sources. An important issue in information integration is the possibility of information conflicts among the different information sources.

## 7.4. Parallelization in context sensitive mining

Context sensitive mining uses textual data as the base for its operations. The enormity of this data requires an efficient processing technique that is both fast and has the capacity to handle the data variations. Hence we can say that a parallel processing environment is best suited for these kind of problems rather than a sequential environment. The parallel processing environments considered for our study are the MapReduce and CUDA architectures.

## 7.5. MapReduce and CUDA: A Comparative Study

MapReduce uses unstructured data for processing, while CUDA handles structured data. In order to process unstructured data in CUDA, it must be structured by the user. Both support parallelization when considering string to vector operations and calculating the word count and comparison. Initially MapReduce worked on Hadoop and CUDA on GPU. Currently, both are interoperable. CUDA can be used in Hadoop and MapReduce can be implemented on GPUs. MapReduce works best in batch processing scenarios, while CUDA works best on both batch processing and in stream analysis. While CUDA performs best in floating point operations, MapReduce works efficiently on string manipulation operations.

## 7.6. Proposed Research Framework for Context Sensitive Mining

The data for Sentiment Analysis can rightly be called Big Data due to various reasons. The high volume nature, the unstructured nature and variety of content present in the data makes it a challenge for the processing environment. Due to this nature, we need to perform a variety of operations in parallel to make it suitable for real time processing of streaming data rather than relying on batch processing which is not suitable for several applications including dynamic content filtering for security.

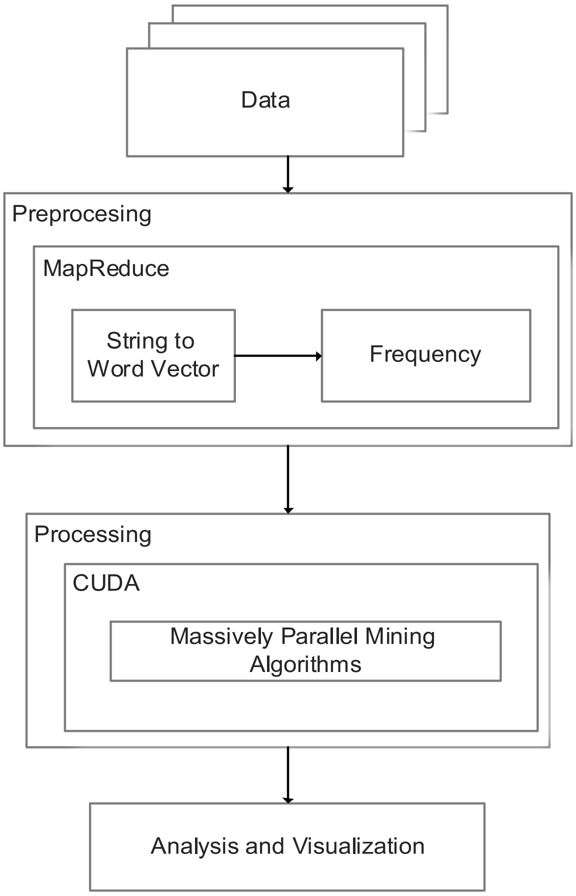


**Fig 5: Context Sensitive Sentiment Analysis Framework**

Here, we describe a framework that provides a direction in which we can bring about higher accuracy along with better efficiency. The framework can be divided into three distinct phases; the pre-processing phase, processing phase and the analysis and visualization phase.

The Big Data that is to be used for analysis is textual in nature. This proves to be the base data for our system. Hence we use MapReduce. MapReduce contains string based processing mechanisms can work efficiently in such cases. Hence it proves to be the best choice for the pre processing technique. In perspective of context sensitive mining pre-processing refers to analyzing the given text and producing vector strings and frequency mapping. This converts the textual data into numerical type, for the next phase of analysis.

The next and the major phase of the framework is the processing phase. This phase processes the numerical data generated by the pre-processing phase using various algorithms to return the results. Since numerical data is involved here, CUDA proves to be the best approach for floating point operations, or in general numerical operations. Hence the processing phase, when performed using CUDA architecture can provide best results.

The analysis and visualization phase is used to examine the results and provide appropriate interpretations for the result along with effective visual representations. GPGPU based CUDA architecture is much suitable for the visual interpretation of the results in real time. Big Data visualization is yet another domain which provides a lot of research directions.

## 8. CONCLUSION

This paper discussed in detail some of the applications from the social media including Facebook, Twitter, Pinterest, LinkedIn and Google+. This paper presented the research issues in handling the information from these social media applications and also the privacy concerns when information from these applications are fused to form a semantically higher level representation of data for cognitive processing. This paper discussed the need for multilingual semantics in such applications that transcends various geographical and cultural regions. This paper also discussed the various research issues in handling and analysing the various types of data formats supported by these applications and how they can be used for several other applications from targeted marketing and content sharing to homeland security. We considered the effectiveness of CUDA and Map-Reduce based massive parallelism for effective real time processing of this information and how they could be combined to provide better results in a lesser time frame. A framework for pre-processing, processing and visualisation of big data based on a hybrid GPGPU Map-Reduce based environment. This paper also discussed the pros and cons of each and also the effectiveness of going for a hybrid approach and proposed a framework for performing efficient sentiment analysis.

## 9. REFERENCES

[1] Shen, Xuehua, Bin Tan, and ChengXiangZhai. 2005.Context-sensitive information retrieval using implicit feedback. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.

[2] Gracia, Jorge, et al. 2012.Challenges for the multilingual Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web 11- 63-71.

[3] Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.

[4] Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics.

[5] Loia, Vincenzo, and Sabrina Senatore. 2013. A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content. Knowledge-Based Systems.

[6] Ghazi, Diman, Diana Inkpen, and Stan Szpakowicz. 2014. Prior and contextual emotion of words in sentential context. *Computer Speech & Language* 28.1: 76-92.

[7] Li, Yung-Ming, Chun-Te Wu, and Cheng-Yang Lai. 2013. A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship.*Decision Support Systems* 55.3:740-752.

[8] Wang, Gang, et al. 2014. Sentiment classification: The contribution of ensemble learning. Decision Support Systems 57: 77-93.

[9] Wijnhoven, Fons, and Oscar Bloemen. 2013.External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews?. Decision Support Systems.

[10] Lei, Jingsheng, et al. 2014.Towards building a social emotion detection system for online news. *Future Generation Computer Systems*.

[11] Mohammad, Saif M. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems* 53.4: 730-741.

[12] Bradbury, Danny. 2011. Data mining with Linkedin. Computer Fraud & Security2011.10: 5-8.

[13] Montejo-Ráez, Arturo, et al. 2014.Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech & Language* 28.1: 93-107.

[14] Lee, Anthony JT, et al. 2013.Discovering content-based behavioral roles in social networks. *Decision Support Systems*.

[15] Khan, Farhan Hassan, Saba Bashir, and UsmanQamar. 2013. TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems.

[16] Nettleton, David F. 2013.Data mining of social networks represented as graphs. Computer Science Review 7: 1-34.

[17] Ortigosa, Alvaro, Rosa M. Carro, and José Ignacio Quiroga. 2014. Predicting user personality by mining social interactions in Facebook. Journal of Computer and System Sciences 80.1: 57-71.

[18] Upadhyaya, Sujatha R. 2013.Parallel approaches to machine learning—A comprehensive survey. Journal of Parallel and Distributed Computing 73.3: 284-292.

[19] Jung, Jason J. 2012.Online named entity recognition method for microtexts in social networking services: A case study of twitter. Expert Systems with Applications39.9: 8066-8070.

[20] Zhou, Jingyu, Yunlong Zhang, and Jia Cheng. 2014. Preference-based mining of top-K influential nodes in social networks. Future Generation Computer Systems31: 40-47.

[21] Kajdanowicz, Tomasz, PrzemyslawKazienko, and WojciechIndyk. 2014. Parallel processing of large graphs. *Future Generation Computer Systems* 32: 324-337.

[22] KamelBoulos, Maged N., et al. 2010.Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine 100.1: 16-23.

[23] Jang, Haeng-Jin, et al. 2013. Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. Expert Systems with Applications 40.18: 7492-7503.

[24] Huang, Jing, et al. 2013. Decentralized mining social network communities with agents. Mathematical and Computer Modelling 57.11: 2998-3008.

[25] He, Qing, et al. 2011.A parallel incremental extreme SVM classifier.Neurocomputing 74.16: 2532-2540.

[26] Lobachev, Oleg, Michael Guthe, and Rita Loogen. 2013. Estimating parallel performance. Journal of Parallel and Distributed Computing 73.6: 876-887.

[27] http://snap.stanford.edu/data/index.html

[28] http://www.nvidia.in/object/cuda_home_new.html

[29] Karloff, Howard, SiddharthSuri, and Sergei Vassilvitskii. 2010. A model of computation for MapReduce. Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics.

[30] Fadika, Zacharia, et al. 2013.MARIANE: Using MApReduce in HPC environments.Future Generation Computer Systems.

[31] www.wikipedia.org