**A**
**Dissertation**
**On**

# "Sentiment Analysis for Twitter"

**Submitted in Partial fulfillment of the requirement**
**For the award of Degree of**

**MASTER OF TECHNOLOGY**
**Computer Technology and Application**
**Delhi Technological University, Delhi**

**SUBMITTED BY**

**TEEJA MARY SEBASTIAN**
**17/CTA/2K10**

**Under the Guidance of**

**Dr. AKSHI KUMAR**

**Assistant Professor**
**Department of Computer Engineering**
**Delhi Technological University**


**DEPARTMENT OF COMPUTER ENGINEERING**
**DELHI TECHNOLOGICAL UNIVERSITY**
**2012**

# Certificate

This is to certify that the work contained in this dissertation entitled "**Sentiment Analysis for Twitter**" submitted in the partial fulfillment, for the award for the degree of M.Tech in Computer Technology and Applications at **DELHI TECHNOLOGICAL UNIVERSITY** by **TEEJA MARY SEBASTIAN, Roll No. 17/CTA/2K10,** is carried out by her under my supervision. This matter embodied in this project work has not been submitted earlier for the award of any degree or diploma in any university/institution to the best of our knowledge and belief.

**(Dr. AKSHI KUMAR)**
**Project Guide**
**Assistant Professor**
**Department of Computer Engineering**
**Delhi Technological University**

# Acknowledgement

I thank the almighty god and my parents, who are the most graceful and merciful, for their blessing that contributed to the successful completion of this project.

I take this opportunity to express a deep sense of gratitude towards my guide **Dr. AKSHI KUMAR**, for providing excellent guidance, encouragement and inspiration throughout the project work. Without her invaluable guidance, this work would never have been a successful one.

I would also like to thank all my classmates for their valuable suggestions and helpful discussions.

**TEEJA MARY SEBASTIAN**

**(17/CTA/2K10)**

# Abstract

The proliferation of Web-enabled devices, including desktops, laptops, tablets, and mobile phones, enables people to communicate, participate and collaborate with each other in various Web communities, viz., forums, social networks, blogs. Simultaneously, the enormous amount of heterogeneous data that is generated by the users of these communities, offers an unprecedented opportunity to create and employ theories & technologies that search and retrieve relevant data from the huge quantity of information available and mine for opinions thereafter. Consequently, Sentiment Analysis which automatically extracts and analyses the subjectivities and sentiments (or polarities) in written text has emerged as an active area of research.

With the rise of social networking age, there has been a surge of user generated content. Microblogging sites have millions of people sharing their thoughts daily because of its characteristic short and simple manner of expression. We propose and investigate a paradigm to mine the sentiment from a popular real-time microblogging service, Twitter, where users post real time reactions to and opinions about "everything".

In this thesis, we expound a hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation of the opinion words in tweets. A case study is presented to illustrate the use and effectiveness of the proposed system.

# List of Figures

# List of Tables

# Contents

# Chapter 1 Introduction and Outline

This chapter briefly introduces the research work proposed in the thesis. Section 1.1 gives an overview of the research undertaken. Section 1.2 sets out the research objectives. Section 1.3 illustrates the proposed framework and the main contributions arising from the work undertaken. Section 1.4 presents an outline of this thesis describing the organization of the remaining chapters. Finally, Section 1.5 gives the summary of the chapter.

## 1.1. Introduction

With the accelerating interest in social networking, blogging, and other information-sharing sites brought about by Web [1], more and more time is being spent by people online daily. With the rise of Web 2 .0 [2] applications such as microblogging, forums and social networks, there came reviews, comments, recommendations, ratings and feedbacks generated by users. The user generated content can be about almost anything including politicians, products, people, events, etc. The explosion of user generated content, made it necessary for companies, politicians, service providers, social psychologists, analysts and researchers to mine and analyze the content for different uses. The bulk of this user generated content cannot be handled manually and thus requires the use of automated techniques for mining and analyzing.

Researchers have been interested in automatically detecting sentiment in texts for many years now. The question was, that if texts were labeled as positive or negative and fed into a computer, was there an algorithm that could learn the characteristics of each emotion? The influential work done early on in Sentiment Analysis [3] used movie reviews to train an algorithm that detects sentiment in text. Movie reviews are a good source for this kind of work because they clearly express an opinion, and because they are accompanied by a numeric rating that makes it easier to train learning algorithms on this data.

This thesis focuses on a different Web 2.0 application: microblogging sites-more specifically Twitter [4]. The posts on Twitter or tweets convey information, which reflects the

mood of the Twitosphere or the world of twitter. Twitter has become a blend of different types of people - ordinary individuals, celebrities, politicians, companies, activists, etc. Almost all the major news outlets, companies, politicians and celebrities have Twitter account where they post news for their followers. People with Twitter accounts can reply to or retweet the posts . People express their sentiment along with what they are posting, retweeting or replying to.

Many recent uprisings in Tunisia, Egypt and closer home, the Anna Hazare led Anti-Corruption campaign in India, have definitely had social media contributing from beginning to end. Both Facebook [5] and Twitter have had multiplying effect throughout the uprisings. The sentiment carried by Facebook or Twitter posts definitely inspired and galvanized people for more action. Due to the increase of hostile and negative communication over social networking sites like Facebook and Twitter, recently the Government of India tried to allay concerns over censorship of these sites where Web users continued to speak out against any proposed restriction on posting of content. As reported in one of the Indian national newspaper [6] "Union Minister for Communications and Information Minister, Kapil Sibal, proposed content screening & censorship of social networks like Twitter and Facebook". Instigated by this, the research carried out by us was to use sentiment analysis to gauge the public mood and detect any rising antagonistic or negative feeling on social medias. Although, we firmly believe that censorship is not right path to follow, this recent trend for research for sentiment mining in twitter can be utilized and extended for a gamut of practical applications like government policy making, damage control, etc. With the swift development and people's constantly escalating interest in social networking, blogging, and other information-sharing medium brought about by Web [1], more and more time is being spent by people online daily. With the rise of Web 2 .0[2] applications such as microblogging, forums and social networks, there came reviews, comments, recommendations, ratings and feedbacks generated by users. The user generated content can be about almost anything including politicians, products, people, events, etc. The explosion of user generated content, made it necessary for companies, politicians, service providers, social psychologists, analysts and researchers to mine and analyze the content for different uses. The bulk of this user generated content required the use of automated techniques for mining and analyzing since manual mining and analysis are difficult for such a huge content.

Most of the notations used and issues raised in this section are addressed in more detail in later chapters. The remainder of this chapter sets out the research objectives, describes the main contributions of the research work, and presents an outline of this thesis.

## 1.2. Research Objectives

### Statement of Research Question

***"Can Sentiment Analysis comprehend the opinions on Microblogging sites?"***

In response to the identified need to better exploit the knowledge capital in the form of opinions accumulated on microblogging sites(specifically Twitter), this unifying research question can be broken down into the following four questions, each of which will be addressed by this research:

- How can Sentiment Analysis be realized on Web 2.0?
- What methods are to be investigated for capturing opinions on Twitter?
- How can the Sentimental (semantic) Orientation of tweets be determined?
- Finally, what applications can this research serve?

Consequently, the three main research objectives of the work undertaken are:

i. **Research Objective I** – To seek the convergence of Web 2.0 applications and Sentiment Analysis
ii. **Research Objective II** – To propose a hybrid approach involving dictionary and corpus based approaches to find the sentimental orientation of tweets
iii. **Research Objective III** – To find out the real life applications for sentiment analysis on tweets

The objective of this thesis is to find techniques to automatically determine the sentiment of tweets posted and gauge the public mood. It specifically aims at developing a hybrid model involving both dictionary and corpus based methods.

## 1.3. Proposed Framework

To find the semantic orientation of the opinion words in tweets, we propose a novel hybrid approach involving both corpus-based and dictionary-based techniques. We also consider features like emoticons and capitalization as they have recently become a large part of the cyber language. To uncover the opinion direction, we will first extract the opinion words in the tweets and then find out their orientation, i.e., to decide whether each opinion word reflects a positive sentiment, negative sentiment or a neutral sentiment. In our work, we are considering the opinion words as the combination of the adjectives along with the verbs and adverbs. The corpus-based method is then used to find the semantic orientation of adjectives and the dictionary-based method is employed to find the semantic orientation of verbs and adverbs. The overall tweet sentiment is then calculated using a linear equation which incorporates emotion intensifiers too.

## 1.4. Organization of Thesis

This thesis is structured into 5 chapters followed by references and appendix.
Chapter 1 presents the research problem, research objectives, justifies the need for a, and outlines the main contributions arising from the work undertaken.

Chapter 2 provides the essential background and context for this thesis and provides a complete justification for the research work described in this thesis.

Chapter 3 provides the details of the methodology employed and outlines the Sentiment Analysis System (SentiTweet System) that constitutes the proposed approach of the research.

Chapter 4 describes the experimental results obtained from a tweet illustration. It also presents the analysis to account for the tests performed.

Chapter 5 presents future research avenues and conclusions based on the contributions made by this thesis.

## 1.5. **Chapter Summary**

This chapter has laid the foundations for this thesis. It briefly introduced the research problem, research objectives and the proposed solution framework. A justification for the research problem is outlined, together with an explanation of the research methodology used. The next chapter examines the pertinent literature most relevant to this research.

# Chapter 2 Literature Review

The focus of this chapter is to review the prominent and relevant research that has been undertaken related to the proposed approach. Section 2.1 discusses the evolution of web giving an overview of the opportunities offered and the challenges associated with Web 1.0, Web 2.0 and Web 3.0 marking the clear distinction between them. This is followed by section 2.2 which elaborates on the various services and applications being offered by Web 2.0.

## 2.1. Evolution of Web

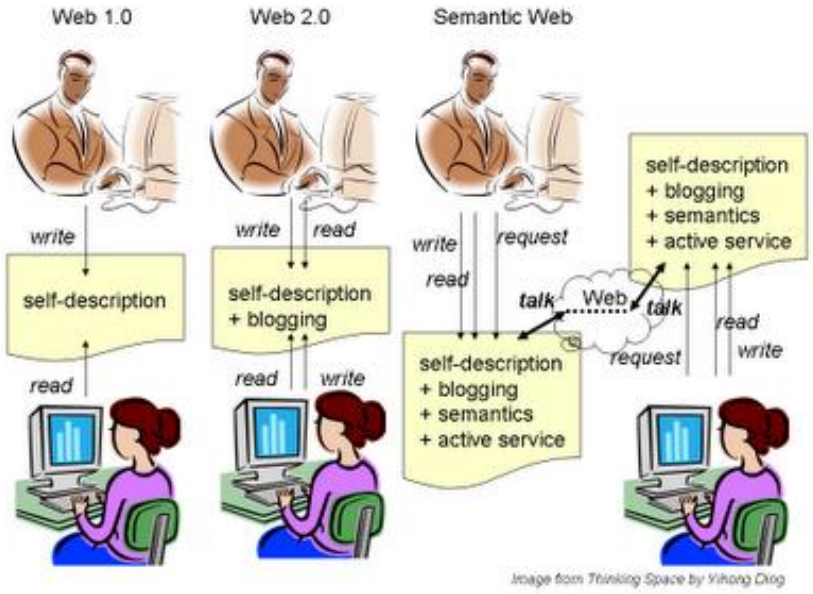
During the last decade, the World Wide Web has evolved into a large worldwide network as announced by many computer experts in the early 1990´s. Many people agree on Web evolution, but few people have thoughtfully studied its principles, i.e. why and how the Web evolves. Web evolution is supposed to be a major branch of Web Science.

One of the early attempts of formalizing the concept of evolution on the Web was done by Tim Berners-Lee in 1989, the father of World Wide Web. In 1998, he explained the importance of evolvability of Web technology. The first concept of Tim Berners-Lee that everyone can contribute was not fulfilled. Although millions of individuals were able to use the Internet, only a small percentage was capable to generate content. The main part of the online users was reading and consuming static pages. In other words, primarily technological aspects dominated the kind of internet access. From this point of view a World Wide Web network is only a connection from static internet pages prepared by some few web developers. The interaction or communication between the typical users was limited by the use of email, discussion forums and also chats. But currently something happens.

The last decade, with the constant inflation of the World Wide Web and the familiarization of the users with the Internet, has generated all the necessary preconditions for a wide adaptation of the basic Internet as a generic exchange platform, where any user becomes a content provider i.e. the kind of web use dramatically changed. This came with the advent of Web 2.0, which is also known as Read/ Write Web. Web 2.0, first coined by Tim O´Reilly in 2004, helps the typical

user to contribute and "The user is the content" is its most popular slogan. The popularity of Web 2.0 grows within all its applications. This new collaborative Web (called Web 2.0), extended by Web-based technologies like comments, blogs and wikis , hosts successful sites like Twitter or Facebook , that allow to build social networks based on professional relationship, interests, etc. The term 'Web 2.0' is defined as the innovative use of the World Wide Web to expand social and business outreach and to exploit collective intelligence from the community. It advocates the Web architecture that promotes users' participation and collaboration and acts as a basic platform for users to share, contribute, review and enhance information resources.

**Figure 2.1 Evolution of Web 2.0**




This picture above shows a simple abstraction of web evolution.

- **Web 1.0 – The World Wide Web**

The traditional World Wide Web, also known as Web 1.0, refers to the original information-oriented web. Web 1.0 is a Read-or-Write Web. In particular, authors of web pages write down what they want to share and then publish it online. Web readers can watch these web pages and subjectively comprehend the meanings. Unless writers willingly release their contact information in their authored web pages, the link between writers and readers is generally disconnected on Web 1.0. By leaving public contact information, however, writers

have to disclose their private identities (such as emails, phone numbers, or mailing addresses). In short, Web 1.0 connects people to a public, shared environment --- World Wide Web. But Web 1.0 does not facilitate direct communication between web readers and writers. In other words, Web 1.0= Websites, E-mail newsletters and "Donate-now" buttons. It is one person or organization pushing content out to many people via websites and e-mail newsletters. It is a one-way communication and the donation process is not interactive or public. One donates and then receives a "Thank You" email.

- **Web 2.0 – The Social Web**

The second stage of web evolution is Web 2.0. The term itself was coined by Dale Dougherty in 2004 and popularized by Tim O'Reilly. It refers to the social web. It's a loose grouping of newer generation social technologies, whose users are actively involved in communicating and collaborating with each other as they build connections and communities across the web [2]. Web 2.0 is a Read/Write Web. At Web 2.0, not only writers but also readers can both read and write to a same web space. This advance allows establishing friendly social communication among web users without obligated disclosure of private identities. Hence it significantly increases the participating interest of web users. Normal web readers (not necessarily being a standard web author simultaneously) then have a handy way of telling their viewpoints without the need of disclosing who they are. The link between web readers and writers becomes generally connected, though many of the specific connections are still anonymous. Whether there is default direction communication between web readers and writers is a fundamental distinction between Web 1.0 and Web 2.0. In short, Web 2.0 not only connects individual users to the web, but also connects these individual users together. It fixes the previous disconnection between web readers and writers. In other words, Web 2.0 = Blogs, Wikis, Social networking sites. It is the beginning of two-way communication in the online public commons. People can post comments and converse with an organization in public for all to see. It's one person or organization publishing content to many on social networking sites who then re-publish the content to their friends, fans, followers, connections, etc. We can also say that, here donation process is a public experience unlike in Web 1.0. Friends, fans, followers, connections, etc. on social

networking sites see the giving and fundraising activity through widgets, apps, and peer-to-peer fundraising tools, like fundraising pages.

- **Web 3.0 - The Semantic Web**

The third stage of web evolution is Web 3.0**.** We don't know precisely what this stage of web evolution is at this moment. It refers to the currently evolving version of the web. There are different conceptions of Web 3.0. Some see Web 3.0 as the semantic web (or the meaning of data), few others see it as a personalization (e.g. iGoogle), and many of them consider it as an intelligent web, where software agents will collate and integrate information to give "intelligent" responses to human operators. This idea is associated with Tim Berners-Lee, the founder of the World Wide Web. Following the last two paradigms, an ideal semantic web is a Read/Write/Request Web. The fundamental change is still at web space. A web space will be no longer a simple web page as on Web 1.0. Neither will a web space still be a Web-2.0-style blog/wiki that facilitates only human communications. Every ideal semantic web space will become a little thinking space. It contains owner-approved machine-processable semantics. Based on these semantics, an ideal semantic web space can actively and proactively execute owner-specified requests by themselves and communicate with other semantic web spaces. By this augmentation, a semantic web space simultaneously is also a living machine agent. We had a name for this type of semantic web spaces as Active Semantic Space (ASpaces). In short, Semantic Web, when it is realized, will connect virtual representatives of real people who use the World Wide Web. It thus, will significantly facilitate the exploration of web resources.

A practical semantic web requires every web user to have a web space by himself . Though it looks abnormal at first glimpse, this requirement is indeed fundamental. It is impossible to imagine that humans still need to perform every request by themselves on a semantic web. Every semantic web space is a little agent. So every semantic web user must have a web space. The emergence of semantic web will eventually eliminate the distinction between readers and writers on the web. Every human web user must simultaneously be a reader, a writer, and a requester; or maybe we should rename them to be web participators. In other words, Web 3.0 = Mobile Websites, Text Campaigns and Smartphone Apps**.** Web 3.0 is all

of the above except that the web experience is no longer limited to desktop and laptop computers while stationary in one place. It's the Internet on the go fueled by mobile phones and tablets. Mobile websites must be designed to be easily read on mobile devices. Group text campaigns function like e-mail newsletters in Web 1.0 to drive traffic to the user's mobile website. Text-to-Give technology allows quick, easy donations on one's mobile phone inspired by urgent calls to actions. Smartphone Apps enable content to be published and shared easily while on the go. Effectively donating via Smartphone Apps doesn't exist yet, but it's coming very soon.

In summary, Web 1.0 connects real people to the World Wide Web. Web 2.0 connects real people who use the World Wide Web. The future semantic web, however, will connect virtual representatives of real people who use the World Wide Web. This is a simple story of web evolution.

| Web 1.0 + Web 2.0 + Web 3.0 = Integrated Web Communications |
|---|

**Note:**

What's important to understand is that all three eras of the web are complimentary and build and serve one another, rather than replace one another. They can also overlap. One uses Web 2.0 tools to drive traffic to the website, to build the e-mail newsletter list, and to increase visits to Donate Now buttons. Also, one uses his Web 2.0 communities to launch Web 3.0 campaigns and uses Web 3.0 tools to grow communities on social networking sites and to send supporters and donors to mobile versions of e-mail newsletter "Subscribe" and "Donate Now" pages. And while many nonprofit communicators are overwhelmed by all these new tools, it's important to understand that there has been a paradigm shift in web communications. Some supporters and donors still prefer to be engaged by your nonprofit Web 1.0 style. Others think "e-mail is for old people" and consistently get most of their content and inspiration from social networking sites. Web 3.0 will organize the masses in ways never seen before through geolocation , group texting and mobile websites, and much of it will be done via Facebook, Twitter, MySpace and Foursquare Smartphone Apps.

**Table 2-1Differences between Web1.0, Web 2.0 and Web 3.0**

| Web 1.0 | Web 2.0 | Web 3.0 |
|---|---|---|
| "The mostly read only web" | "The wildly read-write web" | "The portable personal web" |
| 45 million global users(1996) | 1 billion+ global users(2006) | Focused on individual |
| Focused on companies | Focused on communities | Lifestream |
| Home pages | Blogs | Consolidating dynamic content |
| Owning content | Sharing content | The semantic web |
| Britannica online | Wikipedia | Widgets, drag & drop mashups |
| HTML, portals | XML,RSS | User behavior("me-onomy") |
| Web forms | Web applications | iGoogle, NetVibes |
| Directories("taxonomy") | Tagging("folksonomy") | User engagement |
| Netscape | Google | Advertisement |
| Pages Views | Cost per click | |
| Advertising | Word of mouth | |

**Bottom Line:** There's no "One Fits All" communication tool or tool set anymore. Age, class, race, gender and location play huge roles now in how people want to receive information and calls to action from nonprofits. The good news is that all of these tools are now affordable for nonprofits (even mobile marketing tools!). It's just a matter of keeping up and finding the staff time – and the right person on staff – to master Web 1.0, Web 2.0, and Web 3.0. Those nonprofits that do it best will be the most successful in sharing their mission and programs, creating social change, and securing and maintaining new donors.

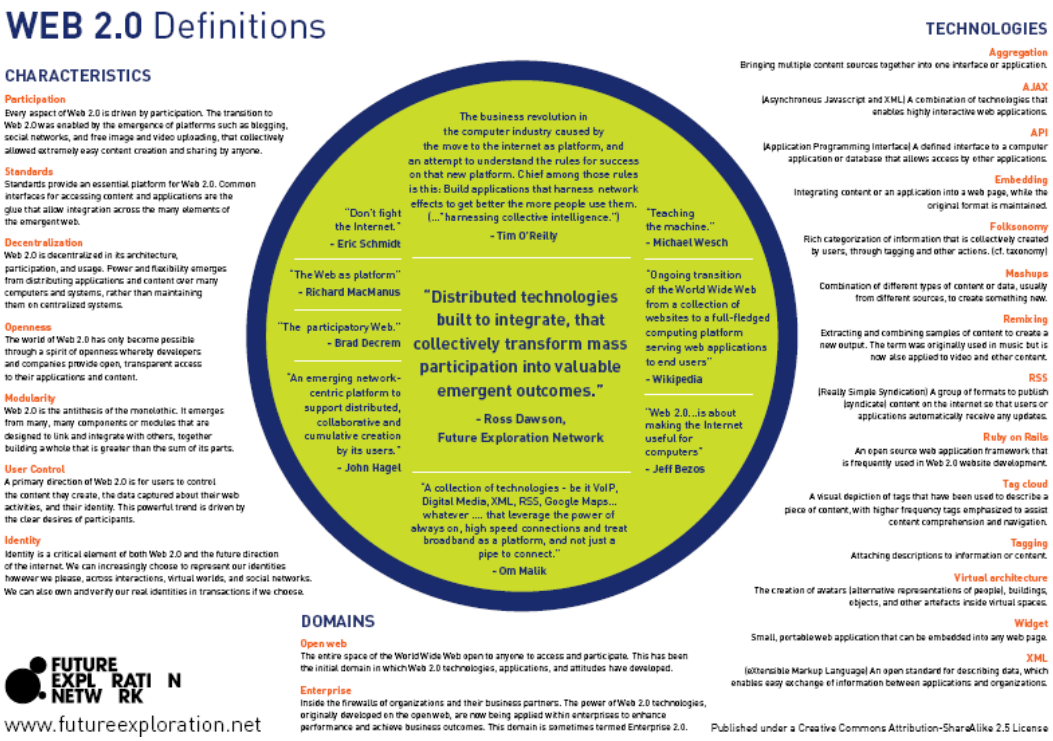
## 2.2. **Web 2.0 services**

The term Web 2.0 was coined on the evolution of web and various web-applications. The term itself was coined by Dale Dougherty in 2004 and popularized by Tim O'Reilly, the founder of World Wide Web. It is not a new concept of web, but a new way of using the web which came after Web 1.0. Although the term suggests a new version of the World Wide Web, it does not

refer to an update to any technical specifications, but rather to cumulative changes in the ways software developers and end-users use the web.

Some popular Web 2.0 tools are podcasting, blogs, RSS, social bookmarking, social networking sites, folksonomies etc. Blogs, wikis and RSS are often held up as exemplary manifestations of Web 2.0. A reader of a blog or a wiki is provided with tools to add a comment or even, in the case of the wiki, to edit the content. This is what we call 'The Read/Write Web'. At Web 2.0, not only writers but also readers can both read and write to a same web space which allows for friendly social communication among web users [7]. What is important to recognize is that the emergence of the Web 2.0 is not a technological revolution, it is a social revolution [8]. This statement means that nowadays the usability of the technology gets simpler and simpler so that we are not forced to learn to use them in a technological way, but in a social way. It is a loose grouping of newer generation social technologies, whose users are actively involved in communicating and collaborating with each other as they build connections and communities across the web. Examples of various social networking sites are myspace.com, friendster.com, facebook.com, multiply.com, tagged.com, twitter.com, etc. Therefore, it is also referred to as '*The Social Web*'. It marks the progression from static web pages to dynamic, interactive ones. A Web 2.0 site gives its users the free choice to interact and collaborate with each other which leads to sharing of information and resources among them. It provides a number of services and applications that facilitate the features such as interactive information sharing, interoperability, and user-centered design. It is most commonly referred to as '*The participatory Web*' that lets its users to actively participate and contribute i.e. there is complete user-involvement. Users can post their comment on news stories , can give their reviews on any information provided online , can upload their photos , can Share digital videos, etc. in contrast to the other websites where users were limited to the passive viewing of content that was created for them. Web 2.0 is '*The User-focused Web*' wherein the user needs are catered as they can freely participate, organize, read, write & play online. Web 2.0 draws together the capabilities of client-side and server-side software, content syndication and the use of network protocols. Standards-oriented web browsers may use plug-ins and software extensions to handle the content and the user interactions. Web 2.0 sites provide users with information storage, creation, and dissemination capabilities that were not possible in the environment now known as "Web 1.0".

**Figure 2.2 Web 2.0 definitions**




Various services of web 2.0 include social-networking sites, blogs, wikis, websites, podcasts, vodcasts, VoIP, RSS, folksonomies and various other web applications.

**Figure 2.3 Web 2.0 Landscape**

**(a) Social-Networking Sites**

Social networking sites with Twitter and Facebook being the best-known - allow users to set up a personal profile page where they can post regular status updates, maintain links to contacts known as 'friends' through a variety of interactive channels, and assemble and display their interests in the form of texts, photos, videos, group memberships, and so on. This might involve drawing in other Web 2.0 tools like RSS feeds, folksonomies, photos, videos, etc, from social sharing sites. Social networking sites represent a fundamental shift from the content-oriented web (where web pages were usually about topics) to the person-oriented web (where web pages are about people).

**(b) Blogs**

Blogs are like online journals where we can post updates - in the form of text, pictures, audio or video files. A blog can function as a reflective diary but it can also be the centerpiece of a conversation, since readers can leave comments for the blog's author and each other, forging connections and community around topics of mutual interest.

**(c) Wikis**

Wikis are collaboratively authored websites, where anyone with a password can make alterations to unlocked sections. The image at left shows a section of the homepage of this wiki as it would appear to a user with editing rights. Wikis rely on the principle of collective intelligence and the notion that the product of collaborative work is often superior to what can be created by a single individual. Advantages for students include the ability to draft and redraft work collaboratively, with each contributor adding to and modifying the work of others. From an educational point of view, wikis are the perfect platform for social constructivist and community of practice approaches, and they are ideal for promoting a sense of a learning community. Feedback can be received from the entire internet (with a public wiki) or class peers (with a private wiki).

The main difference between a weblog and a wiki is that weblogs are personal whereas, wikis are mainly used for collaborative work. For example, if people work on the same documentation, a wiki system should be preferred.

**(d) Podcasts**

Podcasts are audio files, potentially with accompanying text and/or images - though if video is involved, they are referred to not as podcasts but as vodcasts. They are distributed by syndication feeds such as RSS, with each new episode being downloaded to a computer to be played, or else transferred to a mobile device like an iPod or mp3 player. Once we have subscribed to a podcast, new episodes can be received automatically. They can be used, firstly, in a Web 1.0 manner, with teachers recording them and students simply being invited to listen. Such podcasts can range from lecture-style presentations to intensive language learning lessons. They offer many advantages in terms of recycling of material, whether that involves listening to a lecture a second or third time, or listening repeatedly to language learning materials. They have certain advantages over vodcasts as they offer the flexibility to engage in other activities while listening, unlike in vodcasts which require us to watch as well as listen. Podcasts can also be used, secondly, in a more Web 2.0 manner, with students being asked to create their own podcasts for publication to the web.

**(g) Websites**

Websites are made up of web pages (and often include a main page called a homepage). As vehicles for the delivery of information, websites, web pages or homepages have little to do with Web 2.0. In fact, static web pages are one of the most obvious features of Web 1.0. Over time, it has become easier and easier to create such pages. Nowadays, however, there is a whole new generation of websites, web pages and homepages which are dynamic rather than static and have a more Web 2.0 feel and orientation. These often draw in RSS feeds; draw in photos, videos, etc, from social sharing sites and allow user interaction through comments features, discussion boards or chat. For e.g. Flickr is one of the websites which combines a social network with user generated content. Users can work together to collaborate on photo projects and use each others' tags to find new photos. Flickr also has an API for web services to integrate photo collections with blogs and other apps.

## 2.3. **Sentiment Analysis**

A vital part of the information era has been to find out the opinions of other people. In the pre-web era, it was customary for an individual to ask his or her friends and relatives for opinions before making a decision. Organizations conducted opinion polls, surveys to understand the sentiment and opinion of the general public towards its products or services. In the past few years, web documents are receiving great attention as a new medium that describes individual experiences and opinions. With the advent of World Wide Web and specifically with the growth and popularity of Web 2.0 where focus shifted to user generated content, the way people express opinion or their view has changed dramatically. People can now make their opinion, views, sentiment known on their personal websites, blogs, social networking sites, forums and review sites. They are comfortable with going online to get advice. Organizations have evolved and now look at review sites to know how the public has received their product instead of conducting surveys. This information available on the Web is a valuable resource for marketing intelligence, social psychologists and others interested in extracting and mining views, moods and attitude [9].

There is a vast amount of information available on the Web which can assist individuals and organization in decision making processes but at the same time present many challenges as organizations and individuals attempt to analyze and comprehend the collective opinion of others. Unfortunately finding opinion sources, monitoring them and then analyzing them are herculean tasks. It is not possible to manually find opinion sources online, extract sentiments from them and then to express them in a standard format. Thus the need to automate this process arises and sentiment analysis is the answer to this need.

Sentiment analysis or Opinion mining, as it is sometimes called, is one of many areas of computational studies that deal with opinion oriented natural language processing. Such opinion oriented studies include among others, genre distinctions, emotion and mood recognition, ranking, relevance computations, perspectives in text, text source identification and opinion oriented summarization [10]. Sentiment analysis has turned out as an exciting new trend in social media with a gamut of practical applications that range from applications in business(marketing intelligence; product and service bench marking and improvement), applications as sub component technology (recommender systems; summarization; question answering) to applications in politics. It has great potential to be used in business strategies and has helped

organizations get a real-time feedback loop about their marketing strategy or advertisements from the reaction of the public through tweets, posts and blogs. For a new product launch it can give them instant feedback about the reception of the new product. It can gauge what their brand image is, whether they are liked or not.

As the field of sentiment analysis is relatively new, the terminology used to describe this field of research is many. The terms opinion mining, subjectivity analysis, review mining and appraisal extraction are used interchangeably with sentiment analysis. Subjectivity analysis or subjectivity classification is focused on the task of whether the sentence or document is expressing opinions or sentiments of the author or just merely stating facts. Majority of the papers which use the phrase "sentiment analysis" focus on the specific application of classifying reviews as to their polarity (either positive or negative) [10]. The term opinion mining was first noticed in a paper by Dave et al. [11]. The paper defined that an opinion mining tool would "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)". This definition has been broadened to include various other works in this area. The evolution of the phrase sentiment analysis is similar to that of Opinion Mining.

### 2.3.1. **Basic Terminology**

Formally stating Sentiment Analysis is the computational study of opinions, sentiments and emotions expressed in text [3]. The goal of sentiment analysis is to detect subjective information contained in various sources and determine the mind-set of an author towards an issue or the overall disposition of a document.

Wiebe et al. [12] described subjectivity as the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs and speculations. The words opinion, sentiment, view and belief are used interchangeably but there are subtle differences between them [10].

- *Opinion:* A conclusion thought out yet open to dispute ("each expert seemed to have a different opinion").
- *View:* subjective opinion ("very assertive in stating his views").

- *Belief:* deliberate acceptance and intellectual assent ("a firm belief in her party's platform").
- *Sentiment:* a settled opinion reflective of one's feelings ("her feminist sentiments are well-known").

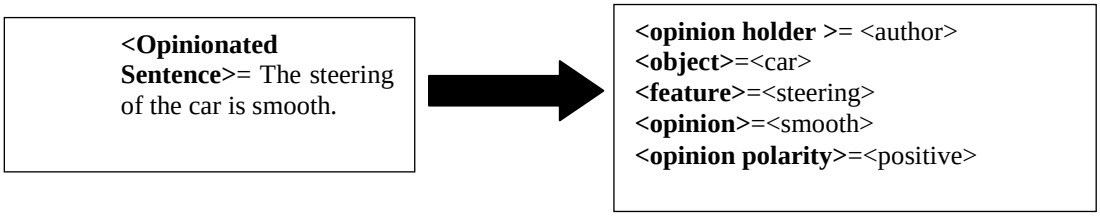
Sentiment analysis is done on user generated content on the Web which contains opinions, sentiments or views. An opinionated document can be a product review, a forum post, a blog or a tweet, that evaluates an object. The opinions indicated can be about anything or anybody, for e.g. products, issues, people, organizations or a service.

Lui[3] mathematically represented an opinion as a quintuple (o, f, so, h, t), where o is an object; f is a feature of the object o; so is the orientation or polarity of the opinion on feature f of object o; h is an opinion holder; t is the time when the opinion is expressed.

- *Object:* An entity which can be a product, person, event, organization, or topic. The object can have attributes, features or components associated with it. Further on the components can have subcomponents and attributes
- *Feature:* An attribute (or a part) of the object with respect to which evaluation is made.
- *Opinion orientation or polarity:* The orientation of an opinion on a feature f indicates whether the opinion is positive, negative or neutral. Most work has been done on binary classification i.e. into positive or negative. But opinions can vary in intensity from very strong to weak[13]. For example a positive sentiment can range from content to happy to ecstatic. Thus, strength of opinion can be scaled and depending on the application the number of levels can be decided.
- *Opinion holder:* The holder of an opinion is the person or organization that expresses the opinion.

  The following example in Figure 2.4 illustrates the basic terminology of sentiment analysis:

**Figure 2.4 Basic Terminology**



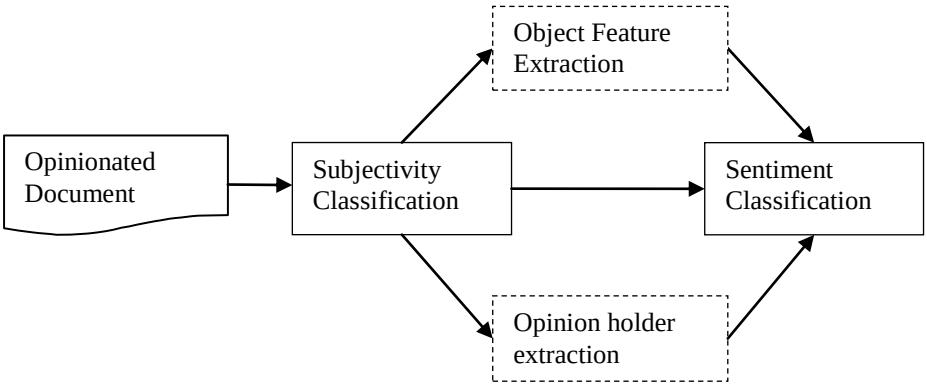
## 2.3.2. **Sentiment Analysis Tasks**

Sentiment analysis is a challenging interdisciplinary task which includes natural language processing, web mining and machine learning. It is a complex task and encompasses several separate tasks, viz:

- Subjectivity Classification
- Sentiment Classification
- Complimentary Tasks
  - ✓ Opinion Holder Extraction
  - ✓ Object Feature Extraction

Figure 2.5 illustrates the major tasks in a sentiment analysis:

**Figure 2.5 Major Tasks of Sentiment Analysis**




The following subsections expound the details of the major tasks in Sentiment Analysis:

### 2.3.2.1. **Subjectivity classification**

Typically, any given document will contain sentences that express opinion and some that do not. That is, a document is a collection of objective sentences, sentences that state a fact, and subjective sentences, sentences that represents the author's opinion, point of view or emotion. Subjectivity classification is the task of classifying sentences as opinionated or not opinionated

[24, 25]. Tang et al. [9], stated subjectivity classification as follows: Let S = {$s_1$, . . . , $s_n$} be a set of sentences in document D. The problem of subjectivity classification is to distinguish sentences used to present opinions and other forms of subjectivity (subjective sentences set $S_s$) from sentences used to objectively present factual information (objective sentences set $S_o$), where $S_s U S_o = S$.

### 2.3.2.2. **Sentiment Classification**

Once the task of finding whether a piece of text is opinionated is over we have to find the polarity of the text i.e., whether it expresses a positive or negative opinion. Sentiment classification can be a binary classification (positive or negative) [14], multi-class classification (extremely negative, negative, neutral, positive or extremely positive), regression or ranking[15]. Depending upon the application of the sentiment analysis, sub -tasks of opinion holder extraction and object feature extraction are optional. (They have been represented by dashed boxes in Figure 2.5)

### 2.3.2.3. **Opinion Holder Extraction**

Sentiment Analysis also involves elective tasks like opinion holder extraction, i.e. the discovery of opinion holders or sources [26, 27]. Detection of opinion holder is to recognize direct or indirect sources of opinion. They are vital in news articles and other formal documents because multiple opinions can be expressed in the same article corresponding to different opinion holders. In documents like these, the multiple opinion holders may explicitly be mentioned by name. In social networks, review sites and blogs the opinion holder is usually the author who may be identified by the login credentials.

### 2.3.2.4. **Object /Feature Extraction**

An additional task is the discovery of the target entity. In contrast with review sites, blogs and social media sites tend not have a set intention or predefined topic and are thus, inclined to discuss assorted topics. In such platforms it becomes necessary to know the target entity.

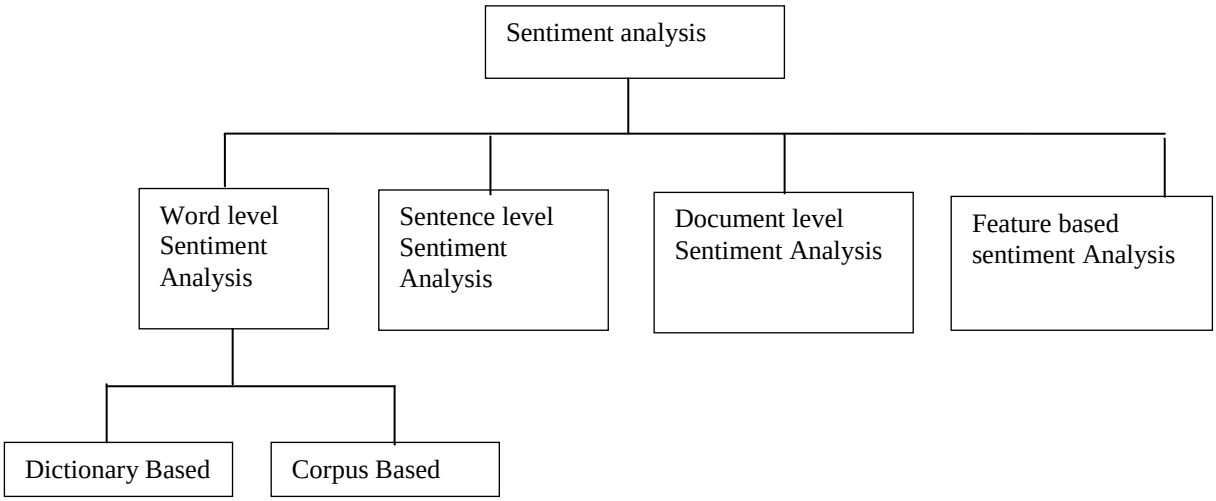
Also as mentioned before target entities can have features or components that are being reviewed. A reviewer can have differing opinions about the different features or components of

the target entity. As a result, feature based sentiment analysis, i.e. extraction of object feature and the related opinion, is an optional task of sentiment analysis [28,29,30].

## 2.3.3. **Levels of Sentiment Analysis**

The tasks described in the previous section can be done at several levels of granularity, namely, word level, phrase or sentence level, document level and feature level. The following figure 4 depicts the levels of granularity of sentiment analysis.

**Figure 2.6 Granularity levels of Sentiment Analysis**




The sentiment analysis tasks can be accomplished at the following levels of granularity:

### 2.3.3.1. **Document Level Sentiment Analysis**

Document-level sentiment analysis considers the whole document as the basic unit whose sentiment orientation is to be determined. To simplify the task, it is presumed that each text's overall opinion is completely held by a single opinion holder and is about a single object. Various machine learning approaches exist for this task. Pang et al. [12] used traditional machine learning methods to classify reviews as positive and negative. They experimented with three classifiers (Naive Bayes, maximum entropy, and support vector machines) and features like unigrams, bigrams, term frequency, term presence and position, and parts-of-speech. They concluded that SVM classifier works best and that unigram presence information was most

effective. Document level sentiment analysis has also been formulated as a regression problem by Pang and Lee [15]. Supervised learning was used to predict rating scores. A simple and straightforward method is to find a linear combination of the polarities in the document, as given by Dave et al. [11] and Turney[31].

The difficulty lies in the fact that there could be mixed opinions in a document, and with the creative nature of natural language, people may express the same opinion in vast ways, sometimes without using any opinion words. Also as stated earlier, a text is equally likely to contain objective sentences along with subjective sentences. Therefore, tools are required to extract useful information from subjective sentences instead of objective ones. This leads to sentence level sentiment analysis.

### 2.3.3.2. **Sentence Level Sentiment Analysis**

At sentence level, research has been done on detection of subjective sentences in a document from a mixture of objective and subjective sentences and then, the sentiment orientation of these subjective sentences is determined. Yu and Hazivassiloglou [32] try to classify subjective sentences and also determine their opinion orientations. For subjective or opinion sentence identification, it uses supervised learning. For sentiment classification of each identified subjective sentence, it used a similar method to Turney[31], but with many more seed words, and log-likelihood ratio as the score function. A simple method used by Liu et al. [33], was to aggregate the orientations of the words in the sentence to get over all polarity of the opinion sentence.

One would expect that subjective sentence detection could be done by using a good sentiment lexicon, but the tricky part is that objective sentences can also contain opinion words

### 2.3.3.3. **Word Level Sentiment Analysis**

The work to find semantic orientation at phrase level is an important task of sentiment analysis. Most works use the prior polarity [34] of words and phrases for sentiment classification at sentence and document levels. Thus, the manual or semi-automatic construction of semantic orientation word lexicon is popular. Word sentiment classification use mostly adjectives as features but adverbs, and some verbs and nouns are also used by researchers [35, 36]. The two

methods of automatically annotating sentiment at the word level are: (1) dictionary-based approaches and (2) corpus-based approaches.

### *Dictionary based Methods*

In this method, a small seed list of words with known prior polarity is created. This seed list is then extended by extracting synonyms or antonyms iteratively from online dictionary sources like WordNet[37]. Kim and Hovy[38] manually created two seed lists consisting of positive and negative verbs and adjectives. They then expanded these lists by extracting, from WordNet, the synonyms and antonyms of the words of the seed list and assigning them to appropriate list (synonyms were placed in the same list and antonyms in the opposite). The sentiment strength of the words was determined by how the new unseen words interacted with the seed list. Both positive and negative sentiment strengths was computed for each word and their relative magnitudes was compared. Based on WordNet lexical relation, Kamps et al. [39] measured the semantic orientation of words.. They collected words and all their synonyms in WordNet, i.e. words of the same synset. Then a graph was created with edges connecting pairs of synonymous words. The semantic orientation of a word was calculated by its relative distance from the two seed terms good and bad. The distance was the length of a shortest path between two words $w_i$ and $w_j$. The values ranged from [-1, 1] with the absolute value indicating the strength of the orientation

The drawback of using a dictionary method is that the polarity classification is not domain specific. For example, "unpredictable" is a positive description for a movie plot but a negative description for a car's steering abilities [31].

### *Corpus based Methods*

Corpus based methods rely on syntactic or statistical techniques like co-occurrence of word with another word whose polarity is known. Hatzivassiloglou and McKeown[40] predicted the orientation of adjectives by assuming that pairs of conjoined adjectives have same orientation (if conjoined by and) and opposite orientation (if conjoined by but). Thus they used conjunctions such as "corrupt and brutal" or "simplistic but well-received" to form clusters of similarly and oppositely-oriented words using a log linear regression model. They intuitively assigned the cluster that contained terms of higher average frequency as the positive list. As this method is an

unsupervised classification method, the corpus required was immense. Turney[31], assigned semantic orientation by using association. That is it is said to have a positive orientation if they have good associations (e.g. Romantic ambience). ). The association relationship between an unknown word and a set of manually-selected seeds (like excellent and poor) was used to classify it as positive or negative The degree of association between the unknown word and the seed words was determined by counting the number of results returned by web searches in the AltaVista Search Engine joining the words with the NEAR operator and calculating the pointwise mutual information between them.

With document, sentence and phrase level analysis, we do not know what the opinion holder is expressing opinion on. Furthermore, we do not know the features that are being talked about.

### 2.3.3.4. **Feature Based Sentiment Analysis**

In a review, its author talks about the positives and negatives of a product. The reviewer may like some features and dislike some, even though the general opinion of the product may be positive or negative. This kind of information is not provided by document level or sentence level sentiment classification. Thus, feature based opinion sentiment analysis [29,30,31] is required. This involves extracting product feature and the corresponding opinion about it. Instinctively, one might think that product features are expressed by nouns and noun phrases, but not all nouns and noun phrases are product features. Yi et al.[28] restricted the candidate words further by extracting only base noun phrases, definite base noun phrase(noun phrases preceded by a definite article “the”) and beginning definite base noun phrases(definite base noun phrase at the beginning of a sentence followed by a verb phrase). For each sentiment phrase detected, its target and final polarity is determined based on a sentiment pattern database. The sentiment pattern database contains sentiment extraction patterns for sentence predicates.

Hu and Lui[33] extract the feature that people are most interested in and thus extract the most frequent noun or noun phrase using association mining. They also extracted infrequent features by extracting the noun or noun phrase nearest to an opinion word in a sentence that contained no frequent features. They use simple heuristic method of assigning the nearest opinion word to a feature to determine the sentiment orientation. Popescu and Etzioni[30] greatly improved the task of extracting features. They distinguish between being a part of an object and a property of the object by using WordNet’s “is-a” hierarchy and morphological clues. Their algorithm tries to

eliminate those noun phrases that probably are not product features. They associated meronymy discriminators with each product class and evaluated noun phrases by computing the PMI (Point-wise Mutual Information) between the phrase and meronymy discriminator.

**Figure 2.7 Sentiment Analysis Model**

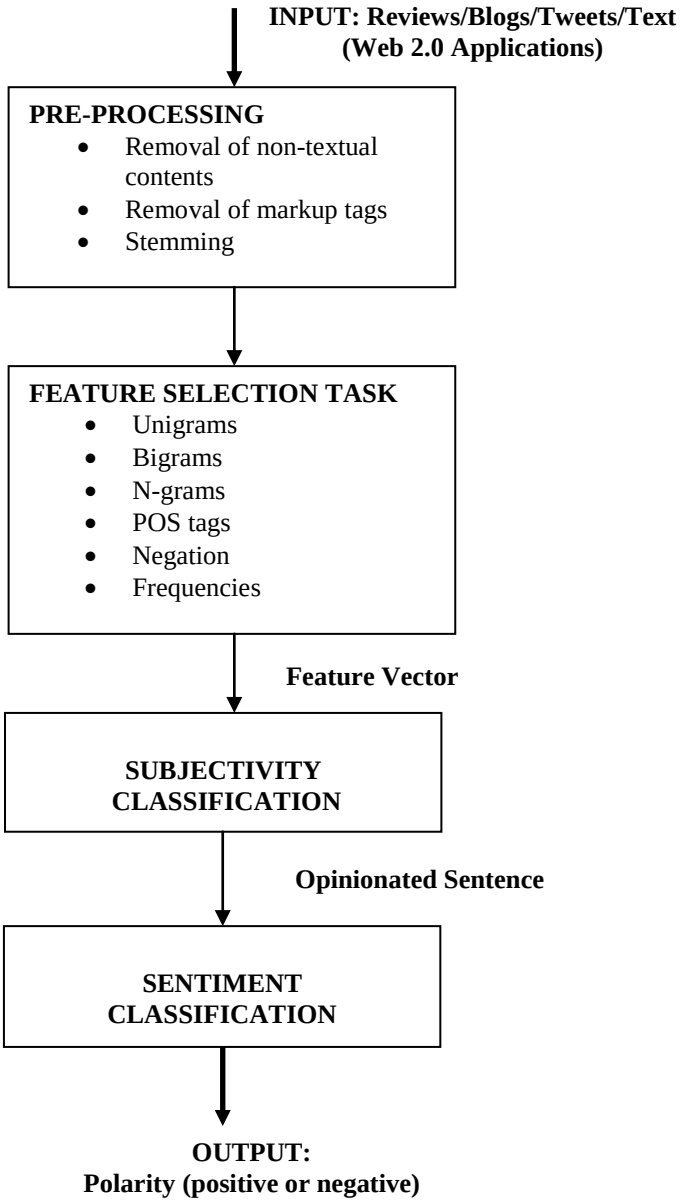


**Table 2-2 Summary of Sentiment Analysis Tasks**

| Sentiment Analysis Tasks |
|---|
| At Document Level<br>• Task: Sentiment Classification of whole document<br>• Classes: positive, negative and neutral<br>• Assumption : Each Document(or review ) focuses on a single object (not true in many discussion posts ) and contain opinion from a single opinion holder |
| At Sentence Level<br>• Task 1: Identifying Subjective/ Objective Sentences<br>  o Classes: Objective and Subjective<br>• Task 2: Sentiment Classification of Sentences<br>  o Classes: positive and negative<br>  o Assumption: A sentence contains only one opinion which may not be true in many cases<br>• *Prior polarities of words determined at word level sentiment analysis is used here* |
| At Feature Level<br>• Task 1: Identify and extract object features that have been commented on by an opinion holder (e.g. A reviewer)<br>• Task 2: Determining whether the opinions on features are negative, positive or neutral<br>• Task 3: Group feature synonyms |

### 2.3.4. **Machine Learning Assisted Sentiment Analysis**

Sentiment Analysis is formulated as a text-classification problem [3] and therefore traditional machine learning techniques are used for the subjectivity/sentiment classification task. High accuracy classification has been achieved by using a variety of techniques, most of which are heavily reliant on machine learning. Like most machine learning applications, the main task of sentiment classification is to engineer a suitable set of features

All text processing approaches require converting text into a feature vector or engineer a suitable set of features. These representations may make the significant features available for machine learning approaches. Few of the features used in practice are given below [10]:

- Words and their frequencies

  Unigrams, bigrams and n-grams along with their frequency counts are considered as features. There has been contention on using word presence rather than frequencies to better describe this feature. Pang et al.[14] so showed better results by using presence instead of frequencies.

- Parts of Speech Tags

  Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment.

- Syntax

  Syntactic patterns like collocations, are used as features to learn subjectivity patterns by researchers. The syntactic dependency patterns can be generated by parsing or dependency trees.

- Opinion Words and Phrases

  Apart from specific words, some phrases and idioms which convey sentiments can be used as features, e.g. ”cost someone an arm and leg”[3].

- Position of Terms

  The position of a term within a text can effect on how much the term affects overall sentiment of the text

- Negation

  Negation is an important but tricky feature to incorporate. The presence of a negation usually changes the polarity of the opinion but all appearances do it. For e.g., “no doubt it is the best in the market” .

As we reviewed the literature for this survey, it was identified that different approaches have been applied to predict the sentiments of words, expressions or documents as to automate the sentiment analysis task. These were either a Natural Language Processing (NLP) research endeavor or addressed by Machine Learning algorithms. Our earlier work [41] probes the role of machine learning as a prominent assisting technology that has ascertained substantial gains in automated sentiment analysis research and practice by developing standards and improving

effectiveness. It expounds the unique aspects of the machine learning techniques in sentiment analysis mainly because of the different features involved in case of supervised[14] and semi-supervised techniques[20,42]. Unsupervised[31] techniques use sentiment driven pattern to obtain labels for words and phrases. While machine learning methods have established to generate good results, there are associated disadvantages. Machine learning classification relies on the training set used, the available literature reports detail classifiers with high accuracy, but they are often tested on only one kind of sentiment source, mostly movie review, thus limiting the performance indication in more general cases. Further, gathering the training set is also arduous; the noisy character of input texts and cross-domain classification add to the complexities and thus push the need for continued development in the area of sentiment analysis.

The research has further substantiated that the existing approaches to sentiment analysis can be grouped into four main categories, namely: *keyword spotting*, where the text is classified in accordance to the presence of reasonably unambiguous affect words; *lexical affinity*, defined as a probabilistic affinity for a particular emotion or opinion polarity to arbitrary words is calculated; *statistical methods*, where the significance of keywords and word co-occurrence frequencies using a large training corpus are computed; and the most recent *sentic computing* [43], based upon a biologically-inspired and psychologically-motivated affective categorization model which makes use of ontologies and common sense reasoning tools for a conceptual-level analysis of natural language text.

We now discuss the various machine learning techniques that have been employed by various researchers over the years for sentiment analysis problem. This section focuses on the unique aspects of the machine learning techniques in sentiment analysis mainly because of the different features involved in case of supervised and semi-supervised techniques. Unsupervised techniques use sentiment driven pattern to obtain labels for words and phrases.

### 2.3.4.1. **Supervised**

Supervised learning generally functions as follows: in the initial training phase, an inductive process learns the characteristics of a class based on a feature set of pre classified documents (reference corpus) and it then applies the acquired knowledge to categorize unseen documents, during testing. Several classical classifiers like Naïve Bayes,, Maximum Entropy and Support Vector Machines are most commonly used supervised methods used.

Pang et al.[14] experimented with three classifiers(Naive Bayes, maximum entropy, and support vector machines) using features like unigrams, bigrams, term frequency, term presence and position, and Parts-of-speech to classify movie reviews as good or bad. They concluded that SVM classifier works best and that unigram presence information was most effective. Dave et al. [11] although claim that in some situations, bigrams and trigrams produce better product-review polarity classification.

Using supervised learning for predicting the rating scores has also been done (1-5 stars) in [15]. The problem is formulated as a regression problem since the rating scores are ordinal.

Supervised learning methods have been used for subjectivity classification too. Most works focus on adjectives and their effects on subjectivity of sentences [44]. Wiebe et al.[45] used the naive Bayes classifier to develop a gold standard data set for subjectivity classification.

Yu and Hatzivassiloglou [32] developed three approaches to classify opinions from facts at the sentence level. The first approach explored the hypothesis that "within a given topic, opinion sentences will be more similar to other opinion sentences than to factual sentences". The second method trained a Naive Bayes classifier , using sentences in opinion and fact documents as the examples of the two categories. The features included words, bigrams, and trigrams, as well as the parts of speech in each sentence. They also included in their features the counts of positive and negative words in the sentence , as well as counts of the polarities of sequences of semantically oriented words. Third approach involved training separate Naive Bayes classifier for each different subset of the features. The goal was to reduce the training set to the sentences that are most likely to be correctly labeled. They assumed as ground truth the information provided by the document labels and that all sentences inherit the status of their document as opinions or facts. Then they train the first classifier on the entire training set. Then they used the classifier to predict the labels of the training set. The sentences that were labeled incorrectly were removed. The second classifier then trained on the reduced training set and this went on until the training set could no longer be reduced.

Wilson et al. [34] also formulate sentiment detection as a supervised learning task. However, instead of using just text classification, they focus on the construction of linguistic features, and train classifiers using Boostexter [46]. Incorporating background knowledge, in terms of linguistic rules, in such classifiers is an interesting direction for future work.

There has been a growing interest in the use of background, prior or domain knowledge in supervised learning. Most of this work has focused on using such prior class-bias of features to generate labeled examples that are then used for standard supervised learning. Provided with some features associated with each class, Wu and Srihari [47] assigned labels to unlabeled documents, which were then used in conjunction with labeled examples to build a Weighted Margin Support Vector Machine.

Another paper that includes prior knowledge is [19]. They constructed a generative model based on a lexicon of sentiment-laden words, and a second model trained on labeled documents. The distributions from these two models were then adaptively pooled to create a composite multinomial Naïve Bayes classifier that captured both sources of information. By exploiting prior lexical knowledge they dramatically reduced the amount of training data required. In addition, by using some labeled documents they were able to refine the background knowledge, which is based on a generic lexicon, thus effectively adapting to new domains.

Adaption to different domains is crucial as the accuracy of sentiment classification can be influenced by the domain. Thus, classifiers trained in a certain domain give poor results in other domains. This is because phrases can be expressing different sentiments in different domains.

#### 2.3.4.2. **Unsupervised**

There has been shift from using supervised approaches to using unsupervised and semi supervised approaches as the manual effort to annotate a huge corpus is too much. Unsupervised learning approaches first build a sentiment lexicon in an unsupervised manner, and then resolve the strength of sentiment (or subjectivity) of a text using a function based on the orientation (or subjectivity) indicators.

Thus, an important task of applying this technique is the construction of the lexicon by means of unsupervised labeling of words or phrases with their sentiment orientation or subjectivity status.

To create a lexicon Turney [31] suggested comparing whether a phrase was more likely to co-occur with the word “poor” or “excellent”. The basic idea was that a phrase has a positive semantic orientation when it has good associations and similarly negative semantic orientations when it has bad associations. The relationship between an unknown word and a set of manually-

selected seeds defined by PMI (Point-wise mutual information), was used to place it into a positive or negative subjectivity class.

Kim and Hovy [38] manually created a small seed list of positive and negative words that contained verbs and adjectives. The synonyms and antonyms of the words were extracted from WordNet and then added to appropriate lists (synonyms would have same orientation and antonyms opposite). The seed lists were further developed by using the expanded list to extract another set of words. They then calculate the sentiment strength of the unseen word by determining how it interacts with the sentiment seed list.

Kamps[39] measured similarity of words by using distance between words based on WordNet lexical relation. They collected all words in WordNet, and related words that could be synonymous, i.e. were part of the same synset. A graph was created with edges connecting each pair of synonymous words. The distance between two words $w_i$ and wj was the length of a shortest path between wi and wj . The orientation of a term was determined by its relative distance from the two seed terms good and bad. The values ranged from [-1, 1] with the absolute value indicating the strength of the orientation.

Gamon et al. [36] used the unsupervised learning technique for identification of aspects or features. They presented an unsupervised aspect identification algorithm that employed clustering over sentences with each cluster representing an aspect. Sentence clusters were labeled with the most frequent non-stop word stem in the cluster.

### 2.3.4.3. **Semi Supervised**

Semi Supervised Learning models learn from both tagged and untagged data. The untagged data provides no information about subjectivity or sentiment polarity but they contain information about the joint distribution of the classification features. Bootstrapping is usually the technique used in semi supervised learning Bootstrapping is fundamentally to use the output of an existing initial classifier to produce labeled data, to which a supervised learning algorithm is later applied. This method is also called self-training.

Riloff et al.[48] proposed a bootstrapping process to identify subjective patterns. A bootstrapping process is used that learns linguistically rich extraction patterns for subjective (opinionated) expressions. Two high-precision classifiers, Hp-Subj and Hp-Obj, label unannotated data to automatically create a large training set, which is then given to an extraction pattern learning

algorithm. The learned patterns are then used to identify more subjective sentences. A set of syntactic templates was needed to represent the space of possible extraction patterns.

Co-training is another semi supervised method that has been applied. Jin et al. [49] created disjoint training sets for building two initial classifiers. The bootstrap document was then tagged using each of the trained HMM(Hidden Markov Model) based classifiers. The opinion sentences that were agreed upon by both classifiers were extracted and saved in the database if it was unique. The newly discovered data was randomly split and added to the training set of the two classifiers. This bootstrap process was continued until no new data could be discovered.

Graph based semi supervised technique has been used in the task of rating inference by Goldberg and Zhu[50].

The following table 2-3 depicting some previous prominent attempts to study sentiment analysis.

**Table 2-3 State -of -Art**

| **Author** | **Granularity Level** | **Model** | **Features** | **Data Source** |
|---|---|---|---|---|
| Hatzivassiloglou and McKeown (1997) [40] | Document | Log Linear Regression | Conjunctions and Adjectives | World Street Journal |
| Das and Chen (2001) [51] | Document | Lexicon and grammar rules | Words | Financial News |
| Pang et al. (2002) [14] | Document | NB[1], SVM[2], ME[3] | Unigram, bigram, contextual effect of negation, frequency, position | IMBD (Movie Review) |
| Turney (2002) [31] | Document | PMI-IR[4] | Bigrams | Automobile, bank, movie, travel reviews |

[1] Naïve Bayes
[2] Support Vector Machines
[3] Maximum Entropy
[4] Pointwise Mutual Information and Information Retrieval

| | | | | |
|---|---|---|---|---|
| Morinaga et al. (2002) [52] | Document | Decision tree induction | Characteristic words, co-occurrence words, and phrases | Cellular phones, PDA and internet service providers |
| Yi et al. (2003) [28] | Topic | NLP- pattern based | Feature lexical semantics | Digital camera and music reviews |
| Turney and Littman (2003) [53] | Document | SO-LSA[5], SO-PMI[6], General inquirer | Words and phrases | TASA-ALL corpus(from sources like novel and news articles) |
| Dave et al. (2003) [11] | Document | Scoring, Smoothing, NB, SVM, ME | Unigrams, bigrams and trigrams | Product reviews |
| Pang and Lee (2004) [54] | Document | NB, SVM | Unigram; Sentence level subjectivity summarization based on minimum cuts. | Movie Reviews |
| Kim and Hovy (2004) [38] | Phrase | Probabilistic based | | DUC corpus |
| Gamon (2004) [55] | Document | SVM | | Customer feedback |
| Nigam and Hurst (2004) [56] | Sentence | syntactic rules based chunking | Lexicon of polar phrase and their parts of speech, syntactic pattern | Usenet message board and other online resources |
| Pang and Lee (2005) [15] | Document | SVM, regression, | | Movie Reviews |

[5] Semantic Orientation Latent Semantic Analysis
[6] Semantic Orientation Point wise Mutual Information

| | | | | |
|---|---|---|---|---|
| | | Metric Labeling | | |
| Choi et al.(2005) [26] | Extract opinion holder, emotion and sentiment | CRF[7] and AutoSlog | Automatically learned extraction patterns | MQPA corpus |
| Wilson et al. (2005) [34] | Phrase | BoosTexter | Subjectivity Lexicon | MQPA corpus |
| Hu and Liu (2005) [29] | Product Feature | Opinion word extraction and aggregation enhanced with WordNet | Opinion words opinion sentences | Amazon Cnn.net |
| Airoldi et al. (2005) [57] | Document | Two stage Markov Blanket Classifier | Dependence among words, minimal vocabulary | IMBd, Infonic |
| Aue and Gamon (2005) [36] | Sentence | NB | Stemmed terms, their frequency and weights | Car reviews |
| Popescu and Etzioni (2005) [30] | Phrase | Relaxation Labeling Clustering | Syntactic dependency template, conjunctions and disjunctions WordNet | Amazon Cnn.net |

[7] Conditional Random Field

| | | | | |
|---|---|---|---|---|
| Cesarano, (2006) [58] | Sentence | Template based using a hybrid evaluation method | POS, n-grams | News articles, web blogs |
| K¨onig and Brill (2006) [59] | Document | Pattern based, SVM, Hybrid | | Movie reviews, customer feedback |
| Kennedy and Inkpen (2006) [60] | Document | SVM, term-counting method, a combination of the two | Term frequencies | General Inquirer dictionary, CTRW dictionary & IMBd |
| Thomas et al.(2006) [61] | Sentence | SVM | Reference Classification | 2005 U.S. floor debate in the House of Representatives |
| Kaji and Kitsuregawa (2007) [42] | Phrase | Phrase trees and word co-occurrence, PMI | lexical relationships, word | HTML documents |
| Blitzer et al. (2007) [62] | Document | Structural Corresponde nce Learning | Word frequency and co-occurrence, part of speech | Book, DVD and kitchen appliance product review |
| Godbole et al. (2007) [63] | Word | Lexical (WordNet) | graph distance measurements between words based on relationships of synonymity and | Newspaper, blog post |

| | | | | |
|---|---|---|---|---|
| | | | anonymity, commonality of words | |
| Annett and Kondrak (2009) [64] | Document | lexical (WordNet) & SVM | number of positive/negative adjectives/adverbs, presence, absence or frequency of words, minimum distance from pivot words | Movie review, blog posts |
| Zhou and Chaovalit (2008) [65] | Document | ontology-supported polarity mining | n-grams, words, word senses | Movie reviews |
| Hou and Li (2008) [66] | Sentence | CRF | POS tags, comparative sentence elements | Product reviews, forum discussions; labeled manually and automatically |
| Ferguson et al. (2009) [67] | Phrase | MNB[8] | binary word feature vectors | Financial blog articles |
| Tan et al.(2009) [68] | Document | NB Classifier with feature adaptation using Frequently Co-occurring Entropy | words | Education reviews, stock reviews and computer reviews |
| Wilson et al. | Phrase | boosting, | words, negation, polarity | MPQA Corpus |

[8] Multinomial Naïve Bayes

| | | | | |
|---|---|---|---|---|
| (2009) [59] | | memory-based learning, rule learning, and support vector learning | modification features | |
| Melville et al. (2009) [19] | Document | Bayesian classification with lexicons and training documents | Words | Blog Posts, reviewing software, political blogs, movie reviews |
| Pak and Paroubek (2010) [21] | Sentence | MNB classifier | N-gram and POS-tags as features | Twitter posts |
| Barbosa and Feng (2010) [22] | Sentence | SVM | retweet, hash tags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words | Twitter posts |
| Heerschop (2011) [70] | document | Creates a list of adjectives and scored | POS, n-grams, negation | Text documents |

While machine learning methods have established to generate good results, there are associated disadvantages. Machine learning classification relies on the training set used, the available literature reports detail classifiers with high accuracy, but they are often tested on only one kind of sentiment source, mostly movie review, thus limiting the performance indication in more general cases. Further, gathering the training set is also arduous; the noisy character of input

texts and cross-domain classification add to the complexities and thus push the need for continued development in the area of sentiment analysis

### 2.3.5. Applications of Sentiment Analysis

The boom in the availability of opinionated and emotionally charged data from various review sites, blog, forums and social networks has created a wave of interest in sentiment analysis by both academia and businesses. This is because there are many practical and potential applications of sentiment analysis. Sentiment analysis assists organizations and service providers to know the mindset of their customers and users and to accordingly tailor their products and services to the needs of customers and users. It is also of vital interest for scientists such as social psychologists as it allows them to tap into the psychological thinking and responses of online communities. Following is a brief discussion on the potential applications of sentiment analysis:

#### *Business Applications*

Sentiment analysis is being adopted by many businesses who would like an edge and an insight into the "market sentiment" [51]. Potential applications would be extracting product review, brand tracking, modifying marketing strategies and mining financial news. The activities that are aided by sentiment analysis are:

1) Automatic tracking of combined user opinions and ratings of brands, products and services from review sites [71].
2) Analyzing purchaser inclinations, competitors, and market trends
3) Gauging reaction to company-related events and incidents, like during a new product launch it can give them instant feedback about the reception of the new product. It can gauge what their brand image is, whether they are liked or not.
4) Monitoring crucial issues to avert harmful viral effects, like dealing with consumer complaints that occur in social media and forwarding the complaints to the particular branch that can handle it, before the grievances multiply.

Key challenges identified by researchers for this application include, identifying aspects of product, associating opinions with aspects of product, identifying fake reviews and processing reviews with no canonical forms.

### *Politics*

Sentiment analysis enables tracking of opinion on issues and subjectivity of bloggers in political blogs. Sentiment analysis can help political organization to understand which issues are close to the voter's heart [23]. Thomas et al. [61], try to determine from the transcripts of US Congressional floor debates which speeches support and which are in opposition to proposed legislation. To improve the worth of the information available to voters, the position of public figures, i.e. causes they support or oppose, can also be determined. Mullen and Malouf [27] describe a statistical sentiment analysis method on political discussion group postings to judge whether there is opposing political viewpoint to the original post. Twitter posts have been used to predict election results [73]. Researchers have collectively pointed out some research challenges namely identifying of opinion holder, associated opinion with issues, identifying public figures and legislation.

### *Recommender System*

Recommender systems can benefit by extracting user rating from text. Sentiment analysis can be used as a sub-component technology for recommender systems by not recommending objects that receive negative feedback [74]. Pang et al. [14] classified movie reviews as "recommended" and "not recommended".

### *Expert Finding*

There is potential of using sentiment analysis in expert finding systems. Taboada et al. [75], use sentiment analysis techniques to track literary reputation. Piao et al. [76] resolve if an author is referencing a piece of work for substantiation or as research that he or she disregards.

### *Summarization*

Opinion summarization finds application when the number of online review of a product is large. This may make it hard for both the customer and the product manufactured. The consumer may not be able to read all the reviews and make an informed decision and the manufacturer may not be able to keep track of consumer opinion. Liu et al. [33] thus took a set of reviews on a certain product and (i) identified product features commented on (ii) identified review sentences that give opinions for each feature; and (iii) produced a summary using the discovered

information. Summarization of single documents [54] or multiple documents (multiple viewpoints) [77] is also an application that sentiment analysis can augment.

#### *Government Intelligence*

Government intelligence is one more application for sentiment analysis. It has been proposed by monitoring sources, the increase in antagonistic or hostile communications can be tracked [78]. For efficient rule making, it can be used to assist in automatically analyzing the opinions of people about pending policies or government-regulation proposals. Other applications include tracking the citizen's opinion about a new scheme, predicting the likelihood of the success of a new legislative reform to be introduced and gauging the mood of the public towards a scandal or controversy.

### 2.3.6. **Issues and Challenges of Sentiment Analysis**

Tackling the fuzzy definition of sentiment and the complexity of its expression in text brings up new questions providing abundant opportunities for quantitative and qualitative work. Major challenges are:

#### *Keyword Selection*

Topic based classification usually uses a set of keywords to classify texts in different classes. In sentiment analysis we have to classify the text in to two classes (positive and negative) which are so different from each other. But coming up with a right set of keyword is not a petty task. This is because sentiment can often be expressed in a delicate manner making it tricky to be identified when a term in a sentence or document is considered in isolation. For example, "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut." (Review by Luca Turin and Tania Sanchez of the Givenchy perfume Amarige, in *Perfumes: The Guide*, Viking 2008.) No ostensibly negative words occur [10].

#### *Sentiment is Domain Specific*

Sentiment is domain specific and the meaning of words changes depending on the context they are used in. The phrase "*go read the book*" would be considered favorably in a book

review, but if expressed in a movie review, it suggests that the book is preferred over the movie, and thus have an opposite result [10].

### *Multiple Opinions in a Sentence*

Single sentence can contain multiple opinions along with subjective and factual portions. It is helpful to isolate such clauses. It is also important to estimate the strength of opinions in these clauses so that we can find the overall sentiment in the sentence, e.g., "*The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small for such a great camera*", expresses both positive and negative opinions [10].

### *Negation Handling*

Handling negation can be tricky in sentiment analysis. For example, "*I like this dress*" and "*I don't like this dress*" differ from each other by only one token but consequently are to be assigned to different and opposite classes. Negation words are called polarity reversers and papers [15, 59] have tried to model negation accurately. But there are many complex polarity reversers like "avoid" in *"[it] avoids all cliché's and predictability found in Hollywood movies*" [10] that have to be addressed.

### *Sarcasm*

Sarcasm and irony are very quite difficult to identify. Sarcasm is a very often used in social media.eg *"thank you Janet Jackson for yet another year of Super Bowl classic rock!"* (Twitter). This refers to the supposedly lame music performance in super bowl 2010 and attributes it to the aftermath of the scandalous performance of Janet Jackson in the previous year [79].

### *Implicit Opinion*

Sentiment that appears in text can be characterized as: explicit where the subjective sentence directly conveys an opinion "*We had a wonderful time*", and implicit where the sentence implies an opinion *"The battery lasted for 3 hours"*. Present sentiment analysis models will not be able to detect this implicit opinion as a negative opinion.

### *Comparative Sentences*

A comparative sentence expresses a relation based on similarities or differences of more than one object [3]. Research on classifying a comparative sentence as opinionated or not is limited. Also the order of words in comparative sentences manifests differences in the determination of the opinion orientation. E.g. The sentence, "*Car X is better than Car Y*" communicates a completely opposite opinion from "*Car Y is better than Car X*".

### *Multilingual Sentiment analysis*

Most sentiment analysis research has focused on data in the English language, mainly because of the availability of resources like lexicons and manually labeled corpora. As only 26.8 % of Internet users speak English[9], the construction of resources and tools for subjectivity and sentiment analysis in languages other than English is a growing need. Several methods have been proposed to leverage on the resources and tools available in English by using cross-lingual projections[80].

### *Opinion Spam*

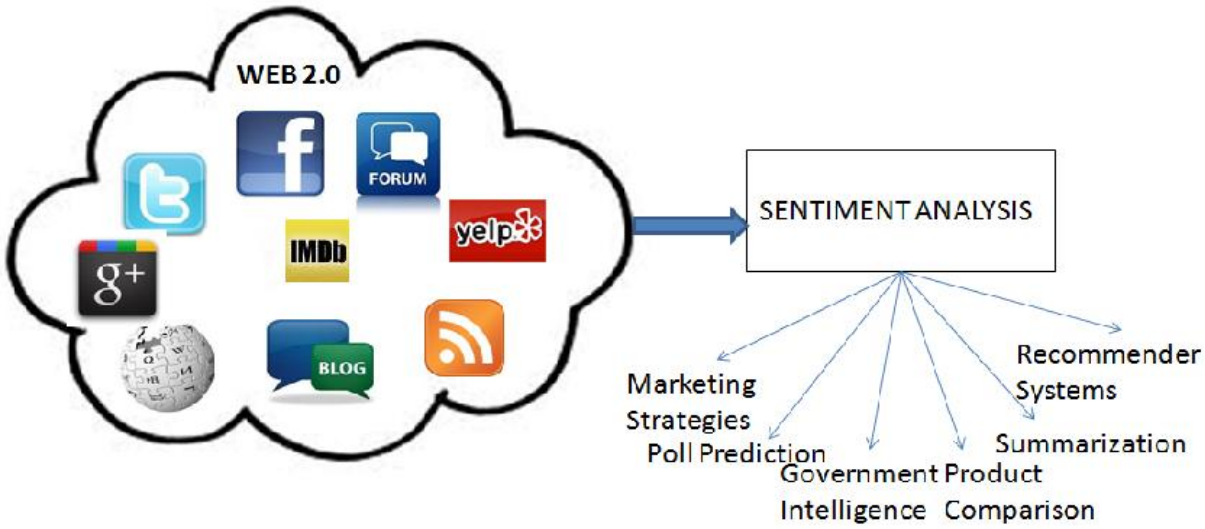
Opinion spam refers to fake or bogus opinions that try to deliberately mislead readers or automated systems by giving undeserving positive opinions to some target objects in order to promote the objects and/or by giving malicious negative opinions to some other objects in order to damage their reputations [3]. Many review aggregation sites try to recognize opinion spam by procuring the helpfulness or utility score of each review from the reader by asking them to provide helpfulness feedbacks to each review ("Was this review helpful?").

## 2.4. Sentiment Analysis and Web 2.0

Web 2.0 is an evolution from passive viewing of information to interactive creation of user generated data by the collaboration of users on the Web. Every aspect of Web 2.0 is driven by participation. The evolution of Web from Web 1.0 to Web 2.0 was enabled by the emergence of platforms the read/ write platforms such as blogging, social networks, and free image and video sharing that collectively allowed extremely easy content creation and sharing by anyone.

[9] http://www.internetworldstats.com/stats7.htm

**Figure 2.8 Conceptual Model of Sentiment Analysis**




The research field of sentiment analysis has been rapidly progressing because of the rich and diverse data provided by Web 2.0 applications. Blogs, review sites, forums, microblogging sites, wikis and social networks have all provided different dimensions to the data used for sentiment analysis.

- ***Review Sites***

  A review site is a website which allows users to post reviews which give a critical opinion about people, businesses, products, or services. Most sentiment analysis work has been done on movie and product review sites[11,14,15]. A review focuses on evaluating a particular object, thus it is a single domain problem. Sentiment analysis on review sites is useful to both manufacturers and potential consumers of the product. The manufacturers can gauge the reception of a product based on the reviews. They can derive the features liked and disliked by the reviewers

- ***Blogs***

  The term web-log or blog, refers to a simple webpage consisting of brief paragraphs of opinion, information, personal diary entries, or links, called posts, arranged chronologically with the most recent first, in the style of an online journal[16]. The bloggers post at hourly, daily or weekly basis which makes the interactions faster and more real time. Desired material within blogs can vary quite widely in content, style, presentation, and even level of

grammaticality. Sentiment analysis on blogs [17,18,19] has been used to predict movie sales, political mood and sales analysis.

- ***Forums***

  Forums or message boards allow its members to hold conversations by posting on the site. Forums are generally dedicated to a topic and thus using forums as a database allows us to do sentiment analysis in a single domain.

- ***Social Networks***

  Social networking is online services or sites which try to emulate social relationships amongst people who know each other or share a common interest. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks. Social network posts can be about anything from the latest phone bought, movie watched, political issues or the individual's state of mind. Thus posts give us a richer and more varied resource of opinions and sentiments.

  - *Twitter*

    Twitter is an online social networking and micro blogging service that enables its users to send and read text-based posts of up to 140 characters, known as "tweets". Sentiment analysis on twitter [20,21,22] is an upcoming trend with it being used to predict poll results[23] among various other applications.

  - *Facebook*

    Facebook is a social networking service and website launched in February 2004. The site allows users to create profiles for themselves, upload photographs and videos. Users can view the profiles of other users who are added as their friends and exchange text messages.

  Social media is the new source of information on the Web. It connects the entire world and thus people can much more easily influence each other. The remarkable increase in the magnitude of information available calls for an automated approach to respond to shifts in sentiment and rising trends

### 2.4.1. Sentiment Analysis and Twitter

It is evident that the advent of real-time information networking sites like Twitter has spawned the creation of an unequaled public collection of opinions about every global entity that is of interest. Although Twitter may provision for an excellent channel for opinion creation and presentation, it poses newer and different challenges and the process is incomplete without adept tools for analyzing those opinions to expedite their consumption.

The area of Sentiment Analysis intends to comprehend these opinions and distribute them into the categories like positive, negative, neutral. Till now most sentiment analysis work has been done on review sites [11]. Review sites provide with the sentiments of products or movies, thus, restricting the domain of application to solely business. Sentiment analysis on Twitter posts is the next step in the field of sentiment analysis, as tweets give us a richer and more varied resource of opinions and sentiments that can be about anything from the latest phone they bought, movie they watched, political issues, religious views or the individuals state of mind. Thus, the foray into Twitter as the corpus allows us to move into different dimensions and diverse applications

#### 2.4.1.1. Data Characteristics

Twitter is a social networking and microblogging service that lets its users post real time messages, called tweets. Tweets have many unique characteristics, which implicates new challenges and shape up the means of carrying sentiment analysis on it as compared to other domains.

Following are some key characteristics of tweets:

- *Message Length:* The maximum length of a Twitter message is 140 characters. This is different from previous sentiment classification research that focused on classifying longer texts, such as product and movie reviews.
- *Writing technique:* The occurrence of incorrect spellings and cyber slang in tweets is more often in comparison with other domains. As the messages are quick and short, people use acronyms, misspell, and use emoticons and other characters that convey special meanings.

- *Availability:* The amount of data available is immense. More people tweet in the public domain as compared to Facebook (as Facebook has many privacy settings) thus making data more readily available. The Twitter API facilitates collection of tweets for training.
- *Topics:* Twitter users post messages about a range of topics unlike other sites which are designed for a specific topic. This differs from a large fraction of past research, which focused on specific domains such as movie reviews.
- *Real time:* Blogs are updated at longer intervals of time as blogs characteristically are longer in nature and writing them takes time. Tweets on the other hand being limited to 140 letters and are updated very often. This gives a more real time feel and represents the first reactions to events.

We now describe some basic terminology related to twitter:

- *Emoticons:* These are pictorial representations of facial expressions using punctuation and letters. The purpose of emoticons is to express the user's mood.
- *Target:* Twitter users make use of the "@" symbol to refer to other users on Twitter. Users are automatically alerted if they have been mentioned in this fashion.
- *Hash tags:* Users use hash tags "#" to mark topics. It is used by Twitter users to make their tweets visible to a greater audience.
- *Special symbols*: "RT" is used to indicate that it is a repeat of someone else's earlier tweet.

Applying sentiment analysis on Twitter is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging. Pak and Paroubek [21] rationale the use microblogging and more particularly Twitter as a corpus for sentiment analysis. They cited:

- Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions.
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.

- Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.
- Twitter's audience is represented by users from many countries.

#### 2.4.1.2. **Related Work**

Parikh and Movassate [81] implemented two Naive Bayes unigram models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model could. Go et al. [20] proposed a solution by using distant supervision, in which their training data consisted of tweets with emoticons. This approach was initially introduced by Read [82]. The emoticons served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. The reported that SVM outperformed other models and that unigram were more effective as features. Pak and Paroubek [21] have done similar work but classify the tweets as objective, positive and negative. In order to collect a corpus of objective posts, they retrieved text messages from Twitter accounts of popular newspapers and magazine, such as "New York Times", "Washington Posts" etc. Their classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features. Barbosa et al. [22] too classified tweets as objective or subjective and then the subjective tweets were classified as positive or negative. They used polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. The feature space used included features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words.

Mining for entity opinions in Twitter, Batra and Rao[83] used a dataset of tweets spanning two months starting from June 2009. The dataset has roughly 60 million tweets. The entity was extracted using the Stanford NER, user tags and URLs were used to augment the entities found. A corpus of 200,000 product reviews that had been labeled as positive or negative was used to train the model. Using this corpus the model computed the probability that a given unigram or

bigram was being used in a positive context and the probability that it was being used in a negative context. Bifet and Frank [84] used Twitter streaming data provided by Firehouse, which gave all messages from every user in real-time. They experimented with three fast incremental methods that were well-suited to deal with data streams: multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They concluded that SGD-based model, used with an appropriate learning rate was the best.

Agarwal et al. [85] approached the task of mining sentiment from twitter, as a 3-way task of classifying sentiment into positive, negative and neutral classes. They experimented with three types of models: unigram model, a feature based model and a tree kernel based model. For the tree kernel based model they designed a new tree representation for tweets. The feature based model that uses 100 features and the unigram model uses over 10,000 features. They concluded features that combine prior polarity of words with their parts-of-speech tags are most important for the classification task. The tree kernel based model outperformed the other two.

# Chapter 3 PROPOSED FRAMEWORK

Chapter 2 identified a number of issues related to Web 2.0 and Sentiment Analysis. This chapter illustrates the novel techniques that constitute the proposed approach to address those issues presented in Chapter 2. Section 3.1 gives an overview of the research undertaken. Section 3.2 depicts the architectural view of the proposed system. Section 3.3 illustrates the proposed system, describes each component of the system and shows how each of the proposed technique contributes to the sentiment analysis process. Finally, Section 3.4 gives the summary of the chapter.

## 3.1. Proposed Framework

The World Wide Web, is a huge, widely distributed, global source for information. It is an ever expanding sea of information. Recent papers reported on the growth of the Web, which by all measures is enormous and growing (in terms of both content and users) at a staggering rate. According to worldwidewebsize.com, the indexed Web contains at least 7.74 billion pages (Tuesday, 22 May, 2012). The advent of real-time information networking sites like Twitter has spawned the creation of an unequaled public collection of opinions about every global entity that is of interest. Although Twitter may provision for an excellent channel for opinion creation and presentation, the process is incomplete without adept tools for analyzing those opinions to expedite their consumption. The field of Sentiment analysis intends to understand these opinions and decompose them into discrete classes like positive, negative, neutral. The foray into Twitter as the corpus allows us to move into different dimensions and diverse applications.

In response to this identified need for knowing others opinions, using twitter as platform, we propose the "SentiTweet system".

To find the semantic orientation of the opinion words in tweets, we propose a novel hybrid approach involving both corpus-based and dictionary-based techniques. We also consider features like emoticons and capitalization as they have recently become a large part of the cyber language. To uncover the opinion direction, we will first extract the opinion words in the tweets and then find out their orientation, i.e., to decide whether each opinion word reflects a positive

sentiment, negative sentiment or a neutral sentiment. In our work, we are considering the opinion words as the combination of the adjectives along with the verbs and adverbs. The corpus-based method is then used to find the semantic orientation of adjectives and the dictionary-based method is employed to find the semantic orientation of verbs and adverbs. The overall tweet sentiment is then calculated using a linear equation which incorporates emotion intensifiers too.

## 3.2. The System Architectural View

The proposed approach intends to accurately retrieve relevant tweets from twitter in response to information need expressed by a user through an inputted query. The tweet is then classified (positive, negative and neutral) and given a score to indicate the intensity. Figure 3.1 shows the architectural overview of the system proposed in this research.

## 3.3. The SentiTweet System

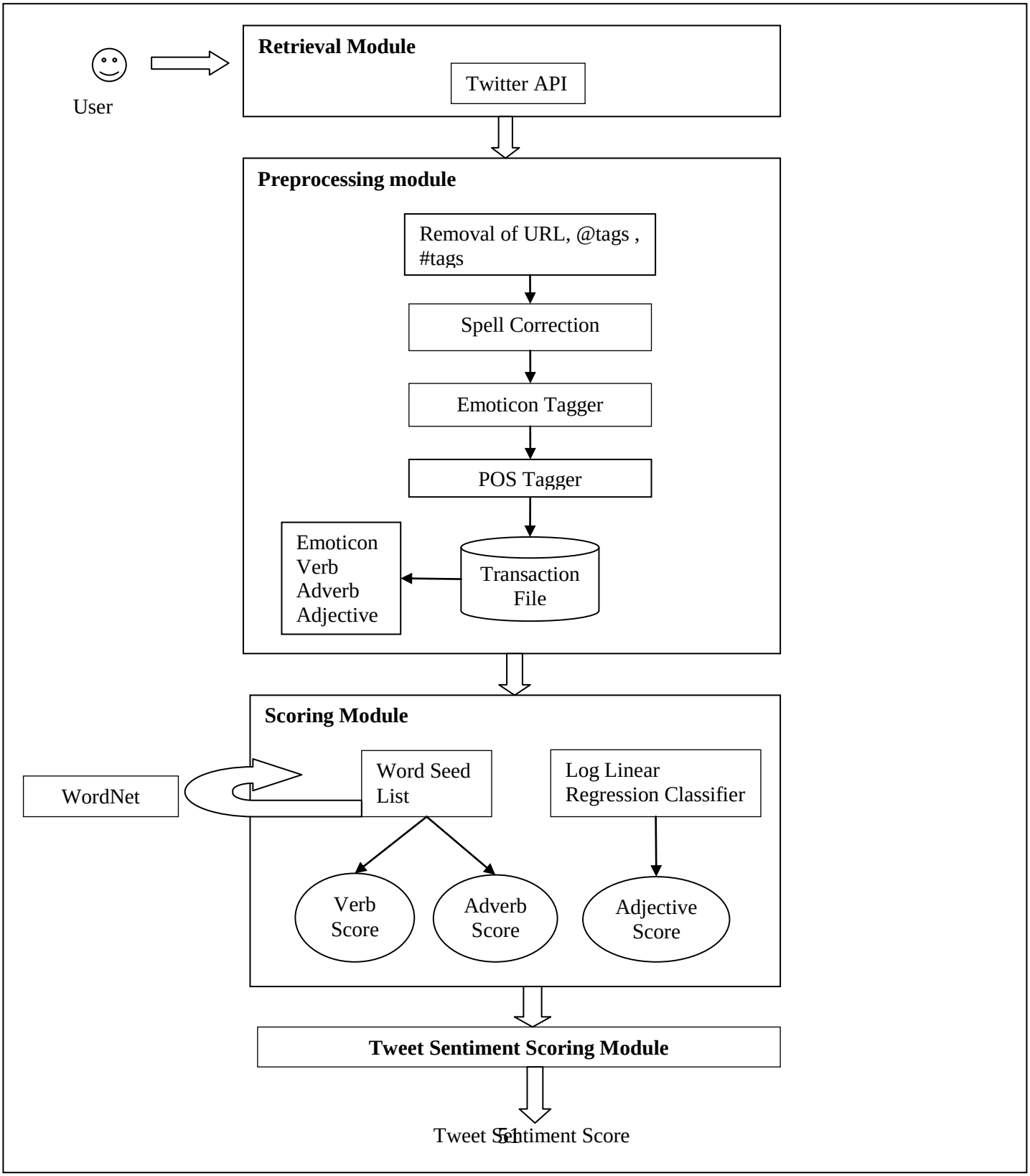
After having examined the principals and objectives of Sentiment Analysis we propose a novel hybrid approach, the SentiTweet system that realizes Sentiment Analysis for Twitter.
The main components of the SentiTweet system are:

- ***Retrieval Module:*** This module uses the Twitter Streaming API to retrieve tweets corresponding to the users query. The tweets retrieved will be about the users query and are the ones that are currently being twittered.
- ***Preprocessing Module:*** This module creates a transaction file to facilitate sentiment analysis. The transaction file consists of adjectives, adverbs, verbs, emoticons and other sentiment intensifiers.
- ***Scoring Module:*** Once we have extracted the adjectives, adverbs, verbs and emoticons, we have to determine their semantic orientation and the intensity of their orientation. For this we score them individually based on dictionary and corpus based methods.
- ***Tweet Scoring Module :*** After finding the individual semantic scores, we have to find out what sentiment the tweet as a whole conveys. We therefore, use a simple equation

that aggregates the scores of the sentiment bearing words and sentiment intensifiers to find the overall sentiment of the tweet.

The following sub-sections expound the details of the SentiTweet system:

**Figure 3.1 SentiTweet System Architecture**



### 3.3.1. **Tweet Retrieval**

Twitter provides us with two APIs: REST and Streaming [86]. The REST API provides simple interfaces for most Twitter functionality. REST API consists of two APIs: one simply called the REST API and another called Search API (their difference is due to their development history). The Streaming API is a family of powerful real-time APIs for Tweets and other social events. It provides near real-time high-volume access to Tweets in sampled and filtered form whereas the REST APIs support short-lived connections and are rate-limited. The Twitter REST API methods allow developers to access core Twitter data. This includes update timelines, status data, and user information. The Search API methods give developers methods to interact with Twitter Search and trends data. REST APIs allow access to Twitter data such as status updates and user info regardless of time. However, Twitter does not make data older than a week or so available. Thus REST access is limited to data Twittered not before more than a week. Therefore, while REST API allows access to these accumulated data, Streaming API enables access to data as it is being Twittered. The Streaming API and the Search REST API can be for data collection. The Streaming API has to used to collect training data is because collecting a large amount of tweets needs a non-rate-limited long-lived connection.

Both the streaming API and the Search REST API have a language parameter that can be set to a language code, e.g. 'en' to collect English data. But the collected data may still contained tweets in other languages making the data very noisy.

### 3.3.2. **Pre-processing of Tweets**

We prepare the transaction file that contains opinion indicators, namely the adjective, adverb and verb along with emoticons (we have taken a sample set of emoticons and manually assigned opinion strength to them). Also we identify some emotion intensifiers, namely, the percentage of the tweet in Caps, the length of repeated sequences & the number of exclamation marks, amongst others. Thus, we pre-process all the tweets as follows:

a) Remove all URLs (e.g. www.example.com), hash tags (e.g. #topic), targets (@username), and special Twitter words (“e.g. RT”).
b) Calculate the percentage of the tweet in Caps.

c) Correct spellings; A sequence of repeated characters is tagged by a weight. We do this to differentiate between the regular usage and emphasized usage of a word.

d) Replace all the emoticons with their sentiment polarity (Table 3-1).

e) Remove all punctuations after counting the number of exclamation marks.

f) Using a POS tagger, the NL Processor linguistic Parser [87], we tag the adjectives, verbs and adverbs.

**Table 3-1 Emoticons**

| ***Emoticon*** | ***Meaning*** | ***Strength*** |
|---|---|---|
| :D | Big grin | 1 |
| BD | Big grin with glasses | 1 |
| XD | Laughing | 1 |
| \m/ | Hi 5 | 1 |
| :),=),:-) | Happy, smile | 0.5 |
| :* | kiss | 0.5 |
| :\| | Straight face | 0 |
| :\ | undecided | 0 |
| :( | sad | -0.5 |
| </3 | Broken heart | -0.5 |
| B( | Sad with glasses | -0.5 |
| :’( | crying | -1 |
| X-( | angry | -1 |

### 3.3.3. **Scoring Module**

The next step is to find the semantic score of the opinion carriers i.e. the adjectives, verbs and adverbs. As mentioned previously, in our approach we use corpus-based method to find the semantic orientation of adjectives and the dictionary-based method to find the semantic orientation of verbs and adverbs.

#### 3.3.3.1. **Semantic Score of Adjectives**

An adjective are a describing word and is used to qualify an object. The semantic orientation of adjectives tend to be domain specific, therefore we use a corpus based approach to quantify the semantic orientation of adjectives in the Twitter domain. Motivated by Hatzivassiloglou and McKeown [40], we ascribe same semantic orientation to conjoined adjectives in most cases and in special cases when the connective is "but", the situation is reversed. Similar to them we apply a log-linear regression model with a linear predictor

$$\eta = w^T x \qquad (1)$$

where x is the vector of observed counts in the various conjunction categories(all *and* pairs, all *but* pairs, all attributive *and* pairs, etc.) for the particular adjective pair and w is the vector of weights to be learnt during training. The response y is non-linearly related to η through the inverse logit function

$$y = \frac{e^{\eta}}{1 + e^{\eta}} \qquad (2)$$

The value y produced denotes the similarity between the words. The seed list of adjectives was taken and assigned semantic scores manually. We also calculated the semantic score of conjoined adjectives by using the manually assigned scores and the similarity value y.

### 3.3.3.2. **Semantic Score of Adverbs and Verbs**

Although, we can compute the sentiment of a certain texts based on the semantic orientation of the adjectives, but including adverbs is imperative. This is primarily because there are some adverbs in linguistics (such as "not") which are very essential to be taken into consideration as they would completely change the meaning of the adjective which may otherwise have conveyed a positive or a negative orientation.

For example;

One user says, "***This is a good book***" and;

Other says, "***This is not a good book***"

Here, if we had not considered the adverb "not", then both the sentences would have given positive review. On the contrary, first sentence gives the positive review and the second sentence gives the negative review. Further, the strength of the sentiment cannot be measured by merely considering adjectives alone as the opinion words. In other words, an adjective cannot alone convey the intensity of the sentiment with respect to the document in question. Therefore, we take into consideration the adverb strength which modify the adjective; in turn modifying the sentiment strength. Adverb strength helps in assessing whether a document gives a *perfect* positive opinion, *strong* positive opinion, a *slight* positive opinion or a *less* positive opinion.

For example;

One user says, ***"This is a very good book"*** and ;

Other says, ***"This is a good book"***

In this example, even though both the users express positive sentiment to the same book, the sentiment intensity they convey is different. In the first sentence, the adverb "very" has further enhanced the adjective "good" , thereby modifying its sentiment strength compared to the second sentence. Therefore, the strength of adverbs is also taken into consideration to accurately measure the sentiment strength.

Some groups of verbs also convey sentiments and opinions (e.g. love, like) and are essential to finding the sentiment strength of the tweet. As adverbs and verbs are not dependent on the domain, we use dictionary methods to calculate their semantic orientation.

In general, words share the same orientations as their synonym and opposite orientations as their antonyms. Using this idea, we propose a simple and effective method by making use of the

synonym set & antonym set in WordNet [Appendix B] to predict the semantic orientation of adverbs and verbs. The seed lists of positive and negative adverbs and verbs whose orientation we know is created and then grown by searching in WordNet [Appendix C]. Based on intuition, we assign the strengths of a few frequently used adverbs and verbs with values ranging from -1 to +1. We consider some of the most frequently used adverbs and verbs along with their strength as given below in Table 3-2.

**Table 3-2 Verb and Adverb Strengths**

| ***Verb*** | ***Strength*** | ***Adverb*** | ***Strength*** |
|---|---|---|---|
| Love | 1 | complete | +1 |
| adore | 0.9 | most | 0.9 |
| like | 0.8 | totally | 0.8 |
| enjoy | 0.7 | extremely | 0.7 |
| smile | 0.6 | too | 0.6 |
| impress | 0.5 | very | 0.4 |
| attract | 0.4 | pretty | 0.3 |
| excite | 0.3 | more | 0.2 |
| relax | 0.2 | much | 0.1 |
| reject | -0.2 | any | -0.2 |
| disgust | -0.3 | quite | -0.3 |
| suffer | -0.4 | little | -0.4 |
| dislike | -0.7 | less | -0.6 |
| detest | -0.8 | not | -0.8 |
| suck | -0.9 | never | -0.9 |
| hate | -1 | hardly | -1 |

Procedure "*determine_orientation"* takes the target Adverb/ Verb whose orientation needs to be determined and the respective seed list as the inputs.

The procedure *determine_orientation* searches Word Net and the Adverb/ Verb seed list for each target adjective to predict its orientation (line 3 to line 8). In line 3, it searches synonym set of the target Adverb/ Verb from the Word Net and checks if any synonym has known orientation from the seed list. If so, the target orientation is set to the same orientation as the synonym (line 4) and the target Adverb/ Verb along with the orientation is inserted into the seed list (line 5). Otherwise, the function continues to search antonym set of the target Adverb/ Verb from the Word Net and checks if any Adverb/ Verb have known orientation from the seed list (line 6). If so, the target orientation is set to the opposite of the antonym (line 7) and the target Adverb/ Verb with its orientation is inserted into the seed list (line 8). If neither synonyms nor antonyms of the target word have known orientation, the function just continues the same process for the next Adverb/ Verb since the word's orientation may be found in a later call of the procedure with an updated seed list.

The complete procedure for predicting adverb and verb polarity is given below:

1. Procedure **determine_orientation** (target_Adverb/ Verb $w_i$ , Adverb/ Verb_ seedlist)
2. begin
3. if ($w_i$ has synonym s in Adverb/ Verb _ seedlist )
4. { $w_i$'s orientation= *s*'s orientation;
5. add $w_i$ with orientation to Adverb/ Verb _ seedlist ; }
6. else if ($w_i$ has antonym a in Adverb/ Verb _ seedlist)
7. { $w_i$'s orientation = opposite orientation of *a*'s orientation;
8. add $w_i$ with orientation to Adverb/ Verb _ seedlist; }
9. end

***Note:***

1) *For those adverbs/ verbs that Word Net cannot recognize, they are discarded as they may not be valid words.*

2) *For those that we cannot find orientations, they will also be removed from the opinion words list and the user will be notified for attention.*
3) *If the user feels that the word is an opinion word and knows its sentiment, he/she can update the seed list.*
4) *For the case that the synonyms/antonyms of an adjective have different known semantic orientations, we use the first found orientation as the orientation for the given adjective.*

### 3.3.4. **Tweet Sentiment Scoring**

As adverbs qualify adjectives and verbs, we group the corresponding adverb and adjective together and call it the adjective group; similarly we group the corresponding verb and adverb together and call it the verb group. The adjective group strength is calculated by the product of adjective score ($adj_i$) and adverb ($adv_i$) score, and the verb group strength as the product of verb score ($vb_i$) and adverb score ($adv_i$). Sometimes, there is no adverb in the opinion group, so the S (adv) is set as a default value 0.5. When there is no adjective or verb in the opinion group, then the S(adj) and S(vb) is set as +1.

Thus,

$$S(AG_i) = S(adj_i) * S(adv_i)$$

$$S(VG_i) = S(vb_i) * S(adv_i)$$

To calculate the overall sentiment of the tweet, we average the strength of all opinion indicators like emoticons, exclamation marks, capitalization, word emphasis, adjective group and verb group as shown below:

$$S(T) = \frac{(1 + (P_c + \log(Ns) + \log(N_x))/3)}{|OI(R)|} * \sum_{i=1}^{|OI(R)|} S(AG_i) + S(VG_i) + N_{ei} * S(E_i) \quad (3)$$

Where,

|OI(R)| denotes the size of the set of opinion groups and emoticons extracted from the tweet,

$P_c$ denotes fraction of tweet in caps,

$N_s$ denotes the count of repeated letters,

$N_x$ denotes the count of exclamation marks,

S ($AG_i$) denotes score of the $i^{th}$ adjective group,

S ($VG_i$) denotes the score of the $i^{th}$ verb group,

S ($E_i$) denotes the score of the $i^{th}$ emoticon

$N_{ei}$ denotes the count of the $i^{th}$ emoticon.

$P_c$, $N_s$ and $N_x$ represent emphasis on the sentiment to be conveyed so they can be collectively called sentiment intensifiers.

If the score of the tweet is more than 1 or less than -1, the score is taken as 1 or -1 respectively.

## 3.4. **Chapter Summary**

This chapter illustrated the proposed SentiTweet system. The next chapter describes the experimental results of along with the analysis for the tests performed to evaluate the novel techniques presented in this chapter.

# Chapter 4 EXPERIMENTAL RESULTS AND ANALYSIS

This chapter describes the experimental results obtained from a document illustration. It also presents the analysis to account for the tests performed

## 4.1. Illustration

To clearly illustrate the effectiveness of the proposed method, a case study is presented with a sample tweet:

<tweet>="@kirinv I hate revision, it's BOOOORING!!! I am totally unprepared for my exam tomorrow :( :( Things are not good...#exams"

### 4.1.1.The pre-processing of Tweet

A transaction file is created which contains the preprocessed opinion indicators.

#### 4.1.1.1. Extracting Opinion Intensifiers

The opinion intensifiers are calculated for the tweet as follows.

1) Fraction of tweet in caps:
   There are a total of 18 words in the sentence out of which one is in all caps. Therefore, $P_c=1/18=0.055$
2) Length of repeated sequence, $N_s=3$
3) Number of Exclamation marks, $N_x=3$

#### 4.1.1.2. Extracting Opinion Words

After the tweet is preprocessed, it is tagged using a POS tagger and the adjective and verb groups are extracted.

The list of Adjective Groups extracted**:**

$AG_1$=totally unprepared

$AG_2$=not good

$AG_3$=boring

The list of Verb Groups extracted:

$VG_1$=hate

The list of Emoticons extracted:

$E_{1 =}$ :(

$N_{e1 =} 2$

## 4.1.2.Scoring Module

Now that we have our adjective group and verb group, we have to find their semantic orientation.

### 4.1.2.1. Score of Adjective Group

S ($AG_1$) = S (totally unprepared) =0.8*-0.5 == -0.4

S ($AG_2$) = S (not good) =-0.8*1= -0.8

S ($AG_3$) = S (boring) = 0.5*-0.25 = -0.125

### 4.1.2.2. Score of Verb Group

S ($VG_1$) = S (hate) = 0.5*-1 = 0.5

## 4.1.3.Tweet Sentiment Scoring

Using the formula defined in equation 3 we can calculate the sentiment strength of the tweet as follows.

$$S(T) = \frac{(1+(P_c+\log(Ns)+\log(N_x))/3)}{|OI(R)|} * \sum_{i=1}^{|OI(R)|} S(AG_i)+S(VG_i)+N_{ei} * S(E_i)$$

$$S(T) = \frac{1.33}{5}\sum_{i=1}^{5} S(AG_i)+S(VG_i)+N_{ei} * S(E_i)$$

$$= \frac{(1.33)}{5} * ((-0.4)+(-0.8)+(-0.125)+(-0.5)+2*(-0.5))$$

$$= -0.751$$

As we have got a negative value, we can classify the tweet as negative. We applied our approach to a sample set of 10 tweets. The semantic analysis results obtained are depicted in table 3 below.

**Table 4-1 Sample Tweets and semantic orientation**

| ***Tweet*** | ***Score*** | ***Orientation*** |
|---|---|---|
| @kirinv I hate revision, it's BOOOORING!!! I am totally unprepared for my exam tomorrow :( :( Things are not good...#exams | -0.751 | Negative |
| Criticism of UID laumched yday is extremely unfair. You may hate or even envy Nilekani but can not deny the idea. | 0.009 | Neutral |
| "@bigDEElight Keeping it real gone wrong, that was hilarious!! And I wonder how often that actually happens IRL! | 0.145 | Positive |
| #iranElection this could get nasty | -0.437 | Negative |
| just getting back from Oaxaca, Mexico by plane | 0.125 | Positive |
| I have created a twitter! This is my ONE AND ONLY twitter guys, someone already stole my url. not too happy about it either :( | -0.24 | Negative |
| Happy happy happy :D | 0.625 | Positive |
| That was pretty much awesome. :) | 0.263 | Positive |
| That other dude sucks!!! | -0.664 | Negative |
| @prncssmojo hey i got a im thingy what is ur screen | 0 | Neutral |

| | | |
|---|---|---|
| name? | | |
| Just got home From work. Dam it wuz tough today | -0.281 | Negative |

# Chapter 5 Conclusion and Future Scope

This chapter draws conclusions based on the contributions made by this thesis, presents the limitations of the study and outlines potential avenues for such investigation in future work.

## 5.1. Research Summary

This thesis set out to solve the practical problem of sentiment analysis of Twitter posts to gauge the public mood. It began with discussion of motivations, theoretical framework and research question, importance of the research, aims and outcomes.

Vast literature of general sentiment analysis, and Twitter-specific sentiment analysis were discussed and possible approaches, techniques, features and assumptions for sentiment analysis of tweets were made.

We then proposed a novel approach that used both dictionary based and corpus based methods to determine the semantic orientation of words. We assigned them sentiment scores to indicate the intensity of their semantic orientation. Finally we found the overall tweet sentiment using a simple equation that incorporated adjectives, adverbs, verbs, emoticons, capitalization, exclamation marks and repetitions.

The major contributions of this research are:

i. We illustrated the convergence of Web 2.0 and Sentiment Analysis. We extensively studied about the various techniques and approaches used by researchers till date for sentiment analysis. We realized that while machine learning methods have established that they generate good results, there are associated disadvantages. Machine learning classification relies on the training set used, the available literature reports detail classifiers with high accuracy, but they are often tested on only one kind of sentiment source, mostly movie review, thus limiting the performance indication in more general cases. Further, gathering the training set is also arduous; the noisy character of input texts and cross-domain classification add to the complexities and thus push the need for continued development in the area of sentiment analysis.

ii. We proposed a novel hybrid approach incorporating dictionary and corpus based methods for finding the sentiment of a tweet. We used the corpus based approach to

determine the orientation and intensity of adjectives. We used the corpus based method for adjectives as adjectives are domain related and thus we could find adjectives specific to the Twitter domain.

## 5.2. Future Research Directions

This study is exploratory in nature and the prototype evaluated is a preliminary prototype. The practice result proves that the proposed system has the characteristics of perceiving the semantic orientation of tweets. The results of this work serve as a partial view of the phenomenon. More research needs to be done in order to validate or invalidate these findings, using larger samples. There are few assumptions, known limitations with respect to the analysis and data and technical challenges in the implemented prototype, which may affect the accuracy of the results.

- We have assumed that the language a tweet is written in is English and this may not be the case. Tweeters tend to use their native languages along with English as it is more close to the language used by them daily.

  E.g. "Gurgaon movie hall zindabad!!" This type of usage may help Twitter users to convey their opinions and feelings in a more natural and comfortable but it becomes exceedingly difficult for learning algorithms to incorporate.
- Another characteristic of tweets is that a lot of acronyms specific to social media is used. Although most of them are objective in nature like "BTW" is "by the way", IDK is "I don't know", "PRT" is "partial retweet", some of them convey sentiment and emotion like "LOL" stands for "laughing out loud" or "FTW" that stands for "For the Win".
- We have not addressed the handling of interrogative sentences. For e.g. "Is the new policy good?". This statement is voicing doubt over a policy in reality but the machine will interpret it as a positive opinion because of the presence of the word good.

If we incorporate the above points, there may be an improvement in the initial results that we have got.

## 5.3. **Conclusion**

The proliferation of microblogging sites like Twitter offers an unprecedented opportunity to create and employ theories & technologies that search and mine for sentiments. The work presented in this paper specifies a novel approach for sentiment analysis on Twitter data. To uncover the sentiment, we extracted the opinion words (a combination of the adjectives along with the verbs and adverbs) in the tweets. The corpus-based method was used to find the semantic orientation of adjectives and the dictionary-based method to find the semantic orientation of verbs and adverbs. The overall tweet sentiment was then calculated using a linear equation which incorporated emotion intensifiers too. This work is exploratory in nature and the prototype evaluated is a preliminary prototype. The initial results show that it is a motivating technique.

# Bibliography

[1] [Online]. Available: World Wide Web Consortium(W3C) www.w3.org/WWW/.

[2] T. O'Reilly, "Web 2.0 Compact Definition: Trying Again," [Online]. Available: http://radar.oreilly.com/archives/2006/12/web_20_compact.html. [Accessed 22 March 2007].

[3] B. Liu, Sentiment Analysis and Subjectivity, in Handbook of Natural Language Processing, Second ed., 2010.

[4] "Twitter," [Online]. Available: http://twitter.com/.

[5] "Facebook," [Online]. Available: www.facebook.com.

[6] [Online]. Available: National Daily, Economic Times: articles.economictimes.indiatimes.com › Collections › Facebook.

[7] Martin Ebner. E-Learning 2.0 = e-Learning 1.0 + Web 2.0. Second International Conference on Availability, Reliability and Security (ARES'07), IEEE.

[8] http://e-language.wikispaces.com/folksonomies

[9] Tang H, Tan S, and Cheng X. A survey on sentiment detection of reviews. Expert Systems with Applications: An International Journal, September 2009, 36(7):10760–10773.

[10] Pang, B and Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieva,l 2008,(1-2),1–135

[11] Dave K., Lawrence S, and Pennock D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of of the 12th international conference on World Wide Web(WWW), 2003, pp.:519–528

[12] Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. Learning subjective language. Computational Linguistics, 2004, 30(3):277–308

[13] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of AAAI, 2004,pages 761–769.

[14] Pang, B., Lee, L., and Vaithyanathan.S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002, (EMNLP):79–86.

[15] Pang B. and Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of the Association for Computational Linguistics (ACL),2005:115–124

[16] Anderson, P. What is Web 2.0? Ideas, technologies and implications for education. Technical report, JISC,2007

[17] Mishne G. and Glance N.Predicting movie sales from blogger sentiment.In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW),2006:155–158.

[18] Liu, Y., Huang, J., An, A., and Yu, X. ARSA: A sentiment-aware model for predicting sales performance using blogs. In Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR),2007

[19] Melville, P., Gryc, W., and Lawrence, R.D. Sentiment analysis of blogs by combining lexical knowledge with text classification.In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.2009, :1275-1284.

[20] Go, A., Bhayani, R., Huang, L. Twitter sentiment classification using distant supervision. Technical report,Stanford Digital Library Technologies Project.2009.

[21] Pak A. and Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, .2010:1320-1326.

[22] Barbosa, L. and Feng, J. Robust Sentiment Detection on Twitter from Biased and Noisy Data.Coling 2010: Poster Volume,36-44.

[23] O'Connor B., Balasubramanyan R., Routledge B.R. , Smith N. A. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. AAAI. 2010

[24] Hatzivassiloglou, V. and Wiebe, J. Effects of adjective orientation and gradability on sentence subjectivity.In Proceedings of the International Conference on Computational Linguistics (COLING) ,2000.

[25] Riloff E. and Wiebe J., Learning extraction patterns for subjective expressions.Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) ,2003.

[26] Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S., Identifying sources of opinions with conditional random fields and extraction patterns.Proceedings of the Human

Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) ,2005.

[27] Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D., Automatic extraction of opinion propositions and their holders. Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004.

[28] Yi, J., Nasukawa, T., Niblack, W., &Bunescu, R.,Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003):427–434

[29] Hu, M. and Liu, B. Mining opinion features in customer reviews. In Proceedings of AAAI,2004: 755–760.

[30] Popescu A-M. and Etzioni O., Extracting product features and opinions from reviews, Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),2005

[31] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL),2005: 417–424.

[32] Yu H. and Hatzivassiloglou V., Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[33] Liu, B., Hu, M., & Cheng, J. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of the 14th international world wide web conference (WWW-2005). ACM Press:10–14.

[34] Wilson T. , Wiebe J., and Hoffmann P.,Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)2005: 347–354

[35] Esuli, A., &Sebastiani, F. Determining the semantic orientation of terms through gloss classification.In Proceedings of CIKM-05, the ACM SIGIR conference on information and knowledge management, Bremen, DE,2005.

[36] Aue, A. and Gamon, M., Customizing sentiment classifiers to new domains: A case study. Proceedings of Recent Advances in Natural Language Processing (RANLP),2005.

[37] "WordNet," [Online], Available: http://wordnet.princeton.edu/

[38] Kim, S. and Hovy, E., Determining the sentiment of opinions.In Proceedings of the International Conference on Computational Linguistics (COLING) ,2004

[39] Kamps, J., Marx, M., Mokken, R.J., de Rijke, M., Using WordNet to measure semantic orientation of adjectives. In Language Resources and Evaluation (LREC),2004.

[40] Hatzivassiloglou, V. and McKeown, K., Predicting the semantic orientation of adjectives.In Proceedings of the Joint ACL/EACL Conference,2004: 174–181

[41] Kumar, A. & Sebastian, T.M., Machine learning assisted Sentiment Analysis. Proceedings of International Conference on Computer Science & Engineering (ICCSE'2012), 123-130, 2012.

[42] Kaji, N. and Kitsuregawa, M.,Building lexicon for sentiment analysis from massive collection of html documents. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2007.

[43] Cambria, Erik. Roelandse, Martijn. ed. Sentic Computing: Techniques, Tools and Applications. Berlin: Springer-Verlag., 2012.

[44] J. Wiebe, "Learning subjective adjectives from corpora," Proceedings of AAAI, 2000

[45] J. Wiebe, R. F. Bruce, and T. P. O'Hara. "Development and use of a gold standard data set for subjectivity classifications." Proceedings of the Association for Computational Linguistics (ACL), pp. 246–253

[46] Schapire, R. E. and Singer, Y. 2000. "BoosTexter: a boosting-based system for text categorization". Machine Learning 39, 2/3, 135–168

[47] X. Wu and R. Srihari." Incorporating prior knowledge with weighted margin support vector machines". In KDD, 2004

[48] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 25– 32, 2003

[49] Wei Jin, Hung Hay Ho, and Rohini K.Srihari.” OpinionMiner: A novel machine learning system for web opinion mining”. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France,2009

[50] Andrew B. Goldberg and Jerry Zhu.” Seeing stars when there aren’t many stars: Graph-based semi-supervised learning for sentiment categorization”. In TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing, 2006

[51] Das, S. and Chen, M., Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA),2001.

[52] Morinaga S., Yamanishi K., Tateishi K., and Fukushima T., Mining product reputations on the web.In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2002: 341–349, Industry track

[53] Turney, P. and Littman, M., Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS),2003, 21(4):315–346.

[54] Pang B. and Lee L., A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (ACL) ,2004: 271–278.

[55] Gamon, M,. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of the International Conference on Computational Linguistics (COLING),2004.

[56] Nigam, K. and Hurst, M.,.Towards a robust metric of opinion.The AAAI Spring Symposium on Exploring Attitude and Affect in Text,2004

[57] Airoldi, E. M., Bai, X., and Padman, R., Markov blankets and meta-heuristic search: sentiment extraction from unstructured text. Lecture Notes in Computer Science,2006, 3932: 167–187.

[58] Cesarano, C., Dorr, B., Picariello, A., Reforgiato, D., Sagoff, A., Subrahmanian, V.: OASYS: An Opinion Analysis System. AAAI Press In: AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (CAAW 2006) :21–26.

[59] K¨onig, A. C. & Brill, E.,.Reducing the human overhead in text categorization. In roceedings of the 12th ACM SIGKDD conference on knowledge discovery and data mining,2006, pp: 598–603.

[60] Kennedy, A. and Inkpen, D., Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2, Special Issue on Sentiment Analysis),2006:110–125.

[61] Thomas M, Pang B., and Lee L., Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2006 :327–335.

[62] Blitzer, J., McDonald, R., and Pereira, F., Domain adaptation with structural correspondence learning.In Empirical Methods in Natural Language Processing (EMNLP), 2006.

[63] Godbole, N., Srinivasaiah, M., and Skiena, S.,Large-scale sentiment analysis for news and blogs.Proceedings of the International Conference in Weblogs and Social Media, 2007.

[64] Annett, M. and Kondrak, G. A comparison of sentiment analysis techniques: Polarizing movie blogs. Advances in Artificial Intelligence,2008, 5032:25–35.

[65] Zhou, L. and Chaovalit, P., Ontology-supported polarity mining.Journal of the American Society for Information Science and Technology,2008, 69:98–110.

[66] Hou, F. and Li, G.-H., Mining chinese comparative sentences by semantic role labeling. Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, 2008.

[67] Ferguson, P., O'Hare, N., Davy, M., Bermingham, A., Tattersall, S., Sheridan, P., Gurrin, C., and Smeaton, A. F., Exploring the use of paragraph-level annotations for sentiment analysis in financial blogs.1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA),2009.

[68] Tan, S., Cheng, Z., Wang, Y., and Xu, H., Adapting naive bayes to domain adaptation for sentiment analysis. Advances in Information Retrieval, 2009, 5478:337–349.

[69] Wilson, T., Wiebe, J., and Hoffmann, P., Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Computational Linguistics, 2009, 35(5):399–433

[70] Heerschop, B., Hogenboom, A., and Frasincar, F. Sentiment lexicon creation from lexical resources, Springer, In 14th International Conference on Business Information Systems (BIS 2011), volume 87 of Lecture Notes in Business Information Processing: 185–196.

[71] Chen Y. Y. and Lee K. V., User-Centered Sentiment Analysis on Customer Product Review. World Applied Sciences Journal 12 (Special Issue on Computer Applications & Knowledge Management),2011: 32-38

[72] Mullen T. and Malouf R.,Taking sides: User classification for informal online political discourse. Internet Research, 2008, 18:177–190.

[73] Tumasjan A., Sprenger T.O., Sandner P.G., Welpe I. M., Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. AAAI ,2010

[74] Terveen L., Hill W., Amento B., McDonald D., and Creter J., PHOAKS: A system for sharing recommendations. In Communications of the Association for Computing Machinery (CACM),2007, 40(3):59–62

[75] Taboada M., Gillies M. A., and McFetridge P., Sentiment classification techniques for tracking literary reputation.In LREC Workshop: Towards Computational Models of Literary Analysis, 2006: 36–43.

[76] Piao S., Ananiadou S., Tsuruoka Y., Sasaki Y., and McNaught J., .Mining opinion polarity relations of citations. In International Workshop on Computational Semantics 84 (IWCS), 2007:366–371. Short paper

[77] Seki Y., Eguchi K., Kando N., and Aono M., Multi-document summarization with subjectivity analysis at DUC 2005. In Proceedings of the Document Understanding Conference (DUC).

[78] Spertus E., Smokey: Automatic recognition of hostile message. In Proceedings of Innovative Applications of Artificial Intelligence (IAAI),1997:1058–1065.

[79] Davidov, D., Tsur, O., and Rappoport, A.,Semi-supervised recognition of sarcastic sentences in twitter and amazon. In Conference on Natural Language Learning(CoNLL), 2010.

[80] Denecke, K.., Using SentiWordNet for Multilingual Sentiment Analysis.Proc.Of the IEEE 24th International Conference on Data Engineering Workshop (ICDEW 2008), IEEE Press:507-512.

[81] R. Parikh and M. Movassate, ”Sentiment Analysis ofUser-Generated Twitter Updates using Various Classification Techniques”, CS224N Final Report, 2009

[82] J. Read. “Using emoticons to reduce dependency in machine learning techniques for sentiment classification”. In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005

[83] S. Batra and D. Rao, ”Entity Based Sentiment Analysis on Twitter”, Stanford University,2010

[84] A. Bifet and E. Frank, ”Sentiment Knowledge Discovery in Twitter Streaming Data”, In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1–15

[85] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data”, In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 , pp. 30–38

[86] “Twitter API,”[Online], Available: https://dev.twitter.com/docs/streaming-apis

[87] “POS Tagger,” [Online], Available: http://www.infogistics.com/textanalysis.html

# APPENDIX A

# POS TAGGING

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context — i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

**NLProcessor - Text Analysis Toolkit**

NLProcessor by Infogistics is a successor for a set of Natural Language Processing technologies developed in the 1990s at the University of Edinburgh. NLProcessor is an engine which handles so-called "low-level" text processing routines: tokenisation, capitalised word normalisation, sentence segmentation, part-of-speech tagging and syntactic chunking which are necessary steps in building many kinds of text handling applications.

NLProcessor outputs linguistic information by directly marking text with XML tags:

- Tokens are represented as "W" elements,
- Word-class Part-Of-Speech information is provided in their "C" attribute,
- Noun and Verb groups are marked as NounGroup and VerbGroup elements, and
- Sentences are marked with "S" elements.

For example,

Consider the following sentence ie. “john has been given 25 bricks” with POS tags:

```
<S>
<NounGroup>
<WC=NNP>John</W>
</NounGroup>
<VerbGroup>
<WC=VBZ>has</W>
<WC=VBN>been</W>
<WC=VBD>given</W>
</VerbGroup>
<NounGroup>
<WC=CD>25</W>
<WC=NNS>bricks</W>
</NounGroup>
<WC=".">.</W>
</S>
```

# APPENDIX B

# WORDNET

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the grouping of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

**Structure**

The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms--words that denote the same concept and are interchangeable in many contexts--are grouped into unordered sets (synsets). Each of WordNet's 117 000 synsets is linked to other synsets by means of a small number of "conceptual relations". Additionally, a synset contains a brief definition ("gloss") and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct meanings are represented in as many distinct synsets. Thus, each form-meaning pair in WordNet is unique.

**Relations**

The most frequently encoded relation among synsets is the super-subordinate relation (also called hyperonymy, hyponymy or ISA relation). It links more general synsets like {furniture, piece_of_furniture} to increasingly specific ones like {bed} and {bunkbed}. Thus, WordNet states that the category furniture includes bed, which in turn includes bunkbed; conversely, concepts like bed and bunkbed make up the category furniture. All noun hierarchies ultimately go up the root node {entity}. Hyponymy relation is transitive: if an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture. WordNet distinguishes among Types (common nouns) and Instances (specific persons, countries and geographic entities). Thus, armchair is a type of chair, Barack Obama is an instance of a president. Instances are always leaf (terminal) nodes in their hierarchies. Meronymy, the part-whole relation holds between synsets like {chair} and {back, backrest}, {seat} and {leg}. Parts are inherited from their superordinates: if a chair has legs, then an armchair has legs as well. Parts are not inherited "upward" as they may be characteristic only of specific kinds of things rather than the class as a whole: chairs and kinds of chairs have legs, but not all kinds of furniture have legs.

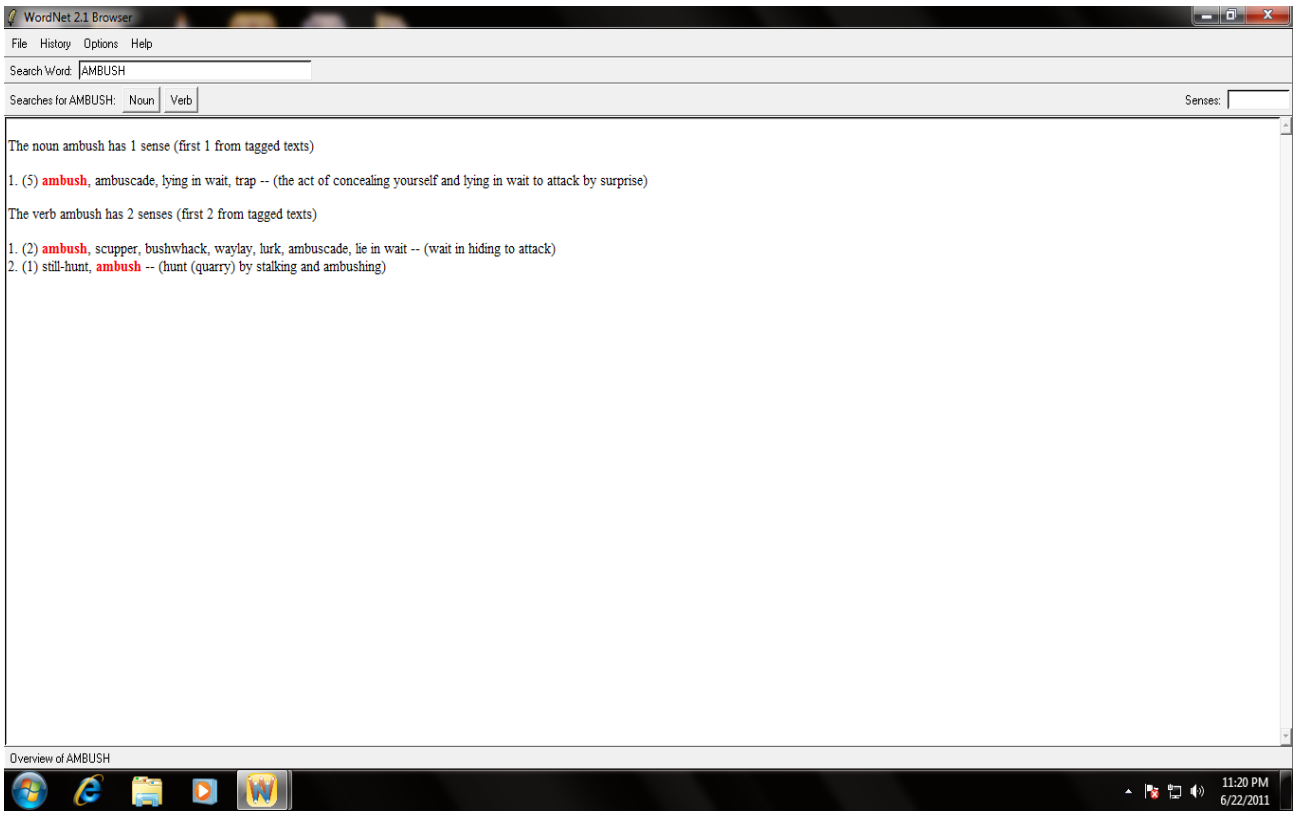
Verb synsets are arranged into hierarchies as well; verbs towards the bottom of the trees (troponyms) express increasingly specific manners characterizing an event, as in {communicate}-{talk}-{whisper}. The specific manner expressed depends on the semantic field; volume (as in the example above) is just one dimension along which verbs can be elaborated. Others are speed (move-jog-run) or intensity of emotion (like-love-idolize). Verbs describing events that necessarily and unidirectionally entail one another are linked: {buy}-{pay}, {succeed}-{try}, {show}-{see}, etc. Adjectives are organized in terms of antonymy. Pairs of "direct" antonyms like wet-dry and young-old reflect the strong semantic contract of their members. Each of these polar adjectives in turn is linked to a number of "semantically similar" ones: dry is linked to parched, arid, dessicated and bone-dry and wet to soggy, waterlogged, etc. Semantically similar adjectives are "indirect antonyms" of the contral member of the opposite pole. Relational adjectives ("pertainyms") point to the nouns they are derived from (criminal-crime). There are only few adverbs in WordNet (hardly, mostly, really, etc.) as the majority of

English adverbs are straightforwardly derived from adjectives via morphological affixation (surprisingly, strangely, etc.).

**Cross-POS relations**

The majority of the WordNet's relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. Cross-POS relations include the "morphosemantic" links that hold among semantically similar words sharing a stem with the same meaning: observe (verb), observant (adjective) observation, observatory (nouns). In many of the noun-verb pairs the semantic role of the noun with respect to the verb has been specified: {sleeper, sleeping_car} is the LOCATION for {sleep} and {painter} is the AGENT of {paint}, while {painting, picture} is its RESULT.

- For example, the figure below is a snapshot of the WordNet 2.1 browser showing the different meanings of the word 'Ambush'.



# APPENDIX C

1. **<tweet>=" Criticism of UID laumched yday is extremely unfair. You may hate or even envy Nilekani but can not deny the idea."**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=0

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=extremely unfair
$AG_2$=even

**The list of Verb Groups extracted:**

$VG_1$=hate
$VG_2$ = not deny

**Emoticons extracted=0**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_1)$=S(extremely unfair)= 0.7*-0.375= - 0.2625
$S(AG_2)$=S(even) = 0.025

**Score of Verb Group**

S ($VG_1$) = S (hate) = 0.5* -1 = -0.5
$S(VG_2)$ =S(not deny)= -0.8 * -0.875=0.7

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{4} * ((-0.5) + (-0.262) + 0.025 + 0.7)$$
$$= -0.009$$

2. **<tweet>="** @bigDEElight Keeping it real gone wrong, that was hilarious!! And I wonder how often that actually happens IRL!**"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=3

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=real
$AG_2$=wrong
$AG_3$=hilarious
$AG_4$=often

**The list of Verb Groups extracted:**

$VG_1$= keeping
$VG_2$ = gone
$VG_3$=wonder
$VG_4$=actually happen

**Emoticons extracted=0**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_{1)}$=S(real)=0.5*0.125= 0.0625
$S(AG_{2)}$=S(wrong)=0.5*-0.75= -0.375
$S(AG_{3)}$=S(hilarious)=0.5*0.375= 0.1875
$S(AG_{4)}$=S(often)=0.25

**Score of Verb Group**

S ($VG_{1)}$=S( keeping )=0.5*0.625=0.3125
$S(VG_{2)}$ = S(gone)= 0.5*-0.375=-0.1875
$S(VG_{3)}$=S(wonder)=0.5*0.375=0.1875
$S(VG_{4)}$=S(actually happen) = 0.375*0.01

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1.477)}{8} * (0.0625 + (-0.375) + 0.1875 + 0.25 + 0.3125 + (-0.1875) + 0.1875 + 0.0375)$$
$$= 0.145$$

3. **<tweet>=“** #iranElection this could get nasty**”**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=0

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=nasty

**Emoticons extracted=0**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_{1)}$=S(nasty)=0.5*-0.875=- 0.437

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{1} * (-0.437)$$
$$= -0.437$$

**4. <tweet>=" just getting back from Oaxaca, Mexico by plane"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=0

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=just

**Emoticons extracted=0**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_{1)}$=S(nasty)=0.125

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{1} * (0.125)$$
$$= 0.125$$

5. **<tweet>="** I have created a twitter! This is my ONE AND ONLY twitter guys, someone already stole my url. not too happy about it either :(**"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=3/24=0.125
Length of repeated sequence=0
Number of Exclamation marks=1

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=not too happy

**The list of Verb Groups extracted:**

$VG_1$= created
$VG_2$ = already stole

**Emoticons extracted=:(**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_{1)}$=S(not too happy)= -0.8*0.6*1= -0.48

**Score of Verb Group**

S ($VG_{1)}$=S( created )=0.5*0.375=0.1875
$S(VG_{2)}$ = S(already stole)= 0.125*-0.5=-0.0625

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1.125)}{4} * ((-0.48) + 0.1875 + (-0.0625))$$
$$= -0.24$$

6. **<tweet>=" Happy happy happy :D"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=0

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=happy
$AG_2$=happy
$AG_{3=}$happy

**Emoticons extracted=:D**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_{1)}$=S(happy)=0.5*1

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{4} * (0.5 + 0.5 + 0.5 + 1)$$
$$= 0.625$$

**7. <tweet>=" That was pretty much awesome :)"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=0

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=pretty much awesome

**Emoticons extracted=:)**

### *Scoring Module*

**Score of Adjective Group**

$S(AG_{1)}$=S(pretty much awesome)=0.3*0.1*0.875=0.0265

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{2} * (0.0265 + 0.5)$$
$$= 0.262$$

8. **<tweet>="** That other dude sucks!!!**"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=3

### *Extracting Opinion Words*

**The list of Verb Groups extracted:**

$VG_1$= sucks

### *Scoring Module*

**Score of Verb Group**

$S(VG_{1)}$=S(sucks)=0.5*-0.9=-0.45

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1.477)}{1} * (-0.45)$$
$$= -0.644$$

9. **<tweet>="** @prncssmojo hey i got a im thingy what is ur screen name?**"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=

### *Extracting Opinion Words*

**The list of Verb Groups extracted:**

None

**The list of Adjective Groups extracted**

None

**Emoticons extracted=0**

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{1} * (0)$$
$$= 0$$

**10. <tweet>="** Just got home From work. Dam it wuz tough today**"**

### *Extracting Opinion Intensifiers*

Fraction of tweet in caps=0
Length of repeated sequence=0
Number of Exclamation marks=0

### *Extracting Opinion Words*

**The list of Adjective Groups extracted:**

$AG_1$=just
$AG_2$=tough

### *Scoring Module*

**Score of Adjective Group**

$S(VG_{1)}$=S(just)=0.125
$S(VG_2)$=S(tough)=0.5*-0.875= -0.4375

### *Tweet Sentiment Scoring*

$$S(T) = \frac{(1)}{2} * (0.125 + (-0.4375))$$
$$= -0.2812$$

# APPENDIX D

## List of Publication

Published

1. Kumar, A. & Sebastian, T.M., (2012), *Machine learning assisted Sentiment Analysis*. Proceedings of International Conference on Computer Science & Engineering (ICCSE'2012), 123-130.

2. Kumar, A. & Sebastian, T.M., (2012), *Sentiment Analysis on Twitter*, International Journal of Computer Science (IJCSI), vol. 9, no. 4, July 2012. (Accepted)

Communicated

1. Kumar, A. & Sebastian, T.M., *Sentiment Analysis: A Perspective on its Past, Present and Future*, (2012) International Journal of Intelligent Systems and Applications (IJISA).