ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 2, Ver. 2 (April - June 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

A Study on Data Sources Exploited in Opinion Mining

[I]Geetika Vashisht, [II]Sangharsh Thakur
[I]Assistant Professor, Kalindi College, Delhi University, India
[II]Lead Engineer, SRI, Noida, India

Abstract
When it's time to reach a decision or choose among myriad of options available for an automobile, gadget, movie, restaurant or any other product or service; people generally prefer to rely on public or peers' experience or opinion. Especially when one has to spend time and money to buy products or services, a prior survey is preferred by all. Until recently, the key sources of information were friends, colleagues, acquaintances, specialized magazine or websites. Now, the World Wide Web (WWW) provides several new tools and platforms to efficiently create and share ideas with everyone. As the communication technology is advancing, people are voicing their opinion online. No one would argue against the power of blogs, forums, review websites, social networks, and content-sharing services in sharing opinions. They give an easy and efficient platform to people for sharing valuable & useful information. Capturing public opinion about social events, political movements, company strategies, marketing campaigns, and product preferences is garnering increasing interest from the scientific community (for the exciting open challenges), and from the business world (for the remarkable marketing fallouts and for possible financial market prediction)[1]. Thus, revolutionizing the need for developing opinion-tracking systems is becoming commercially crucial. Opinion mining is an interesting field focusing on data mining and natural language processing (NLP) techniques to discover, extract, and distill information and opinions from the World Wide Web's vast unstructured textual information. The pre-requisite of opinion mining is to identify the potential data sources for extracting the public's opinion on several products or services. This paper throws a light on the data sources available on the WWW which can contribute to the process of Opinion Mining.

Keywords
Sentiment analysis, Opinion Mining, World Wide Web (WWW), Natural Language Processing/NLP

I. Introduction

Current estimates of the total number of internet users in the world (as on 6th June 2014, 5.30pm) by internet live stats[2] is 2,909,949,700 ; the number of blogs written are 2,546,980; tweets sent are 448,890,789; Facebook active users are 1,258,847,300 and the number is multiplying at a fast pace. This growing trend of consumers turning to blogs, review sites and social media when looking to make a purchase or deciding upon a service has encouraged many companies to use opinion mining and sentiment analysis as part of their research. Marketing and advertising industry deploys systems to continuously gather a wide array of information from the Web, such as products feedback or brands perception which is then used to develop marketing strategies. Other systems might also use opinion mining and sentiment analysis as subcomponent technology to improve customer relationship management and recommendation systems through positive and negative customer feedback [1]. Similarly, opinion mining and sentiment analysis might detect and exclude "flames" (overly heated or antagonistic language) in social communication and enhance anti-spam systems [1]. Designing automatic tools that extracts and synthesize the huge corpus of unstructured data (in the form of reviews, blogs or forums) into valuable structured information is a field of research today. Numerous companies already provide tools that track public viewpoints on a large scale by offering graphical summarizations of trends and opinions in the blogosphere [1]. The remainder of this paper is structured as follows. First, Section 2 throws a light on Opinion Mining and sentiment analysis. In Section 3, we study the categories of data sources like blogs, forums, review sites (paid and non-paid survey sites) and social networking sites. Then in section 4, we identify seven different areas/domains that can be exploited for extracting public opinion for the respective domain. Last, in Section 5, we draw conclusions.

II. Background

Sentiment analysis and Opinion Mining are used interchangeably. Basically, Opinion Mining aims at polarity detection while sentiment analysis focuses on emotion recognition but since the identification of emotion or sentiment is often exploited for detecting polarity, the two fields used as synonyms. Several approaches of sentiment analysis of natural language text have been proposed in the recent years. Opinion tracking systems requires a corpus of data in order to function properly.

According to a survey conducted by Dimensional Research, an overwhelming 90 percent of respondents who recalled reading online reviews claimed that positive online reviews influenced buying decisions, while 86 percent said buying decisions were influenced by negative online reviews [2].

The survey sponsored by Zendesk included responses from 1,046 participants. Approximately two-thirds of the 1,046 respondents reported reading online reviews. While Facebook was the leading resource for positive reviews, the most common place to find negative reviews were online review sites [2].

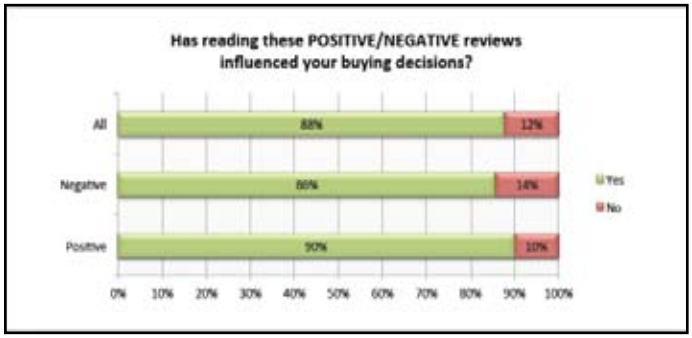

Fig. 1

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

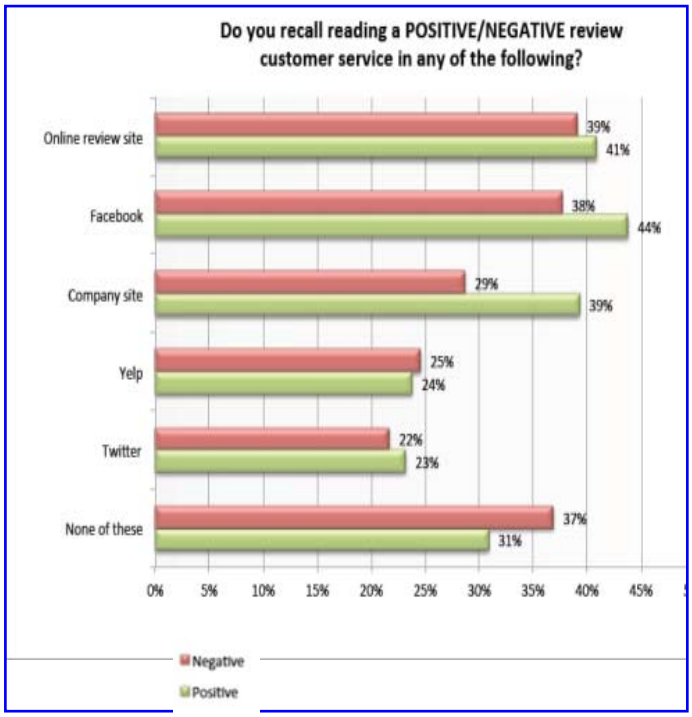

Fig. 2

Not only are customers most frustrated with the way customer service issues are handled, 58 percent said they were more likely to share customer service experiences today than they were five years ago, with more and more people sharing experiences on social networking sites and writing online reviews [2].Of the respondents who shared negative experiences, 45 percent used social media and 35 percent shared via online review sites [2]. Web 2.0 provides great opportunities to express personal experiences and opinions at review sites, forums, discussion groups, blogs etc. In the following sections, we will closely examine the data sources which can be used to feed the opinion tracking systems.

III. Categories of Data Sources

A. Blogs

A blog is a personal website or a web page where people can record opinions on a regular basis. A blog can be generic or restrict itself to discussion of a particular topic. The term "blogging" encompasses the activities like authoring a blog, maintaining a blog or adding any article to an existing blog. A person who posts these entries is called a "blogger". A blog is like an online diary than anybody can read and leave comments on. A blog comprises of text, hypertext, hyperlinks (to other websites, images, videos, and audios), images, and also some emoticons. The blogs differ not only in the type of content, but also in the way that content is delivered or written.

1. Personal blogs

The personal blog is a kind of a diary written by an individual to express his/her feelings on a particular topic.

2. Micro-blogging

Micro-blogging is the type of blog that lets users post small pieces of digital content like text, images, links, short videos, or other media on the Internet. Variety of people uses it to serve different purpose. Friends use it to stay connected to each other, company colleagues use it to coordinate meetings or share useful resources, celebrities and politicians micro-blog about concert dates, events, book releases, or tour schedules. Examples include Twitter, Facebook, Tumblr, and so far the largest WeiBo.

3. Blogs based on category

There is rich profusion of blogs that discuss about gadgets, restaurants, health, beauty services, politics, religion, current events, tourism, fashion, education, music, gardening, quizzes and legal issues, Tutorial and the list goes on.

4. Blogs based on media type

MEDIA TYPE	BLOG'S SPECIFIC TITLE
Videos	Vlogs
Links	Linklogs
Photos	Photoblog
Mixed (short posts + other media type)	Tumblelogs

5. Reverse blog

There is no singe blogger. Many users contribute on a topic by posting their views. Anyone can write to that blog but there is a limit to the number of entries so that it doesn't turn into a forum. Over 409 million people view more than 14.5 billion blogs' pages each month [3]. Users produce about 42.6 million new posts and 54.5 million new comments each month [3]. So, blogs be a great source that can be exploited to infer opinion about a specific product or a service. Several tools that already exist to serve the purpose are [3]

1. SenticNet (http://sentic.net)
2. Luminoso (http://luminoso.com)
3. Factiva (http://dowjones.com/factiva)
4. Attensity (http://attensity.com)
5. Converseon (http://converseon.com)

B. Forums

Nowadays both blogs & forums are arguably the king of self-expression. Both are very popular with their separate methodologies of expression. Forums are usually niche specific. People can voice their opinion and a thread can remain active for years. There are forums that talk about computer technology, about motorcycles, even pets. There is probably a forum for every niche that there is. A forum needs to have a moderator whose primary responsibility is to ensure that everyone follows the rules; especially because the participants can indulge into arguments. If you just want to express yourself in the internet, a blog might be the right tool for you.

C. Review Sites

A review site is a website on which reviews can be posted about people, businesses, products, or services. Review sites are generally supported by advertising. Examples include Amazon, Mouthshut, Qype, Epinions, yelp, Customer Lobby, ConsumerAffairs.com, Judy's Book, Niche, Glassdoor. We can categorize review sites into paid and non-paid survey sites. Paid survey sites pay some amount to the consumers to take part in surveys. Companies get a good amount of feedback about their product or service from these review sites. One of the popular paid review site is www.topconsumerreviews.com/paid-surveys/

D. Social Networking Sites

Opinion Mining & social networks is no doubt a promising match. Social networks provide a perfect solution to the problem of opinion acquisition. Popular social networking sites include Facebook, Twitter, Google+. Table I displays list of top 15 networking sites (in the descending order of their popularity).

Table 1 : [42]

Social Networking Site	Logo	Estimated Unique Monthly Visitors
Facebook		900,000,000
Twitter		310,000,000
LinkedIn		255,000,000
Pinterest		250,000,000
Google+		120,000,000
Tumblr.		110,000,000
Instagram		100,000,000
VK		80,000,000
Flickr		65,000,000
MySpace		42,000,000
Meetup		40,000,000
Tagged		38,000,000
Ask.fm		37,000,000
MeetMe		15,500,000
ClassMates		15,000,000

IV. Domains that are commonly exploited for extracting public opinion

Table 2 : [4-41]

Domain	Forums
Automobile	http://www.autoguide.com/forums http://www.team-bhp.com/forum http://www.automotiveforums.com/ http://gearheads.in/ http://www.theautomotiveindia.com/forums/ http://forums.automotive.com/
Gadgets	http://www.geek.com/forums/forum/gadgets/ http://forum.mobileandgadget.com/
Fashion	http://indiafashionforum.co.in/home/ http://www.fashionbeans.com/forums/ http://forums.thefashionspot.com/ http://www.fashionindustrynetwork.com/forum
Property	http://www.indianrealestateforum.com/ http://www.propertyforum.com/forum/ http://www.magicbricks.com/Real-Estate-Forum http://australianpropertyforum.com/index/

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Movies	http://www.movie-list.com/forum/ http://www.movieforums.com/community/
Health	http://www.healthforum.com/ http://ehealthforum.com/health/health_forums.html http://forums.menshealth.com/
Education	http://entrance-exam.net/forum/ http://www.bhef.com/ http://www.educationforum.co.uk/
Photo-graphy	http://www.dpreview.com/forums http://forum.digitalcamerareview.com/ http://www.thephotoforum.com/forum/ http://photography-on-the.net/forum/ http://www.amateurphotographer.co.uk/forums/
Defence	http://defenceforumindia.com/forum/
Tourism	http://www.tripadvisor.in/ForumHome http://www.travelindiaguide.com/forum/ http://forum.virtualtourist.com/
Sports	http://www.thefootballforum.net/ http://www.indiancricketfans.com/forum.php http://www.totalfootballforums.com/forums/ http://www.basketballforum.com/

V. Conclusions

Now a day, people are increasingly using Internet to express their opinions or emotions. So, it is of utmost importance to identify the data sources that can be fed to opinion tracking systems. In this paper, we have identified and categorised the data sources available on the World Wide Web. We identified the commonly used tools to exploit information present in blogs. Product manufacturers, service providers, marketing & advertising industry rely heavily on the feedback or reviews present on the review sites, blogs and social networks. To achieve the same, opinion tracking systems and tools need to exploit valuable and huge repository of information which is a pre-requisite for opinion mining.

References

[1] *Erik Cambria, National University of Singapore, Björn Schuller, Technical University of Munich, Yunqing Xia, Tsinghua University, Catherine Havasi, Massachusetts Institute of Technology; New Avenues in Opinion Mining and Sentiment Analysis, 2013, IEEE Computer Society.*
[2] *Amy Gesenhues, Third Door Media's General Assignment Correspondent, Marketingland.com*
[3] *www.wordpress.com*
[4] *http://www.thefootballforum.net/*
[5] *http://www.indiancricketfans.com/forum.php*
[6] *http://www.totalfootballforums.com/forums/*
[7] *http://www.basketballforum.com/*
[8] *http://defenceforumindia.com/forum/*
[9] *http://www.dpreview.com/forums/*
[10] *http://forum.digitalcamerareview.com/*
[11] *http://www.thephotoforum.com/forum/forum.php*
[12] *http://photography-on-the.net/forum/*
[13] *http://www.amateurphotographer.co.uk/forums*
[14] *http://www.tripadvisor.in/ForumHome*
[15] *http://www.travelindia-guide.com/forum/default.asp*
[16] *http://forum.virtualtourist.com/*
[17] *http://indiafashionforum.co.in/home/index.html*
[18] *http://www.fashionbeans.com/forums/*
[19] *http://forums.thefashionspot.com/*
[20] *http://www.fashionindustrynetwork.com/forum*
[21] *http://www.movie-list.com/forum/*
[22] *http://www.movieforums.com/community/*
[23] *http://www.autoguide.com/forums.html*
[24] *http://www.team-bhp.com/forum/*
[25] *http://www.automotiveforums.com/*
[26] *http://gearheads.in/*
[27] *http://www.theautomotiveindia.com/forums/*
[28] *http://forums.automotive.com/*
[29] *http://www.geek.com/forums/forum/gadgets/*
[30] *http://forum.mobileandgadget.com/*
[31] *http://www.indianrealestateforum.com/*
[32] *http://www.propertyforum.com/forum/*
[33] *http://www.magicbricks.com/Real-Estate-Forum*
[34] *http://australianpropertyforum.com/index/*
[35] *http://www.healthforum.com/*
[36] *http://ehealthforum.com/*
[37] *http://forums.menshealth.com/*
[38] *http://entrance-exam.net/forum/*
[39] *http://www.bhef.com/*
[40] *http://www.educationforum.co.uk/*
[41] *http://www.topconsumerreviews.com/paid-surveys/*
[42] *http://www.ebizmba.com*