Chapter 5
Web Opinion Mining and Sentimental Analysis

Edison Marrese Taylor, Cristián Rodríguez O., Juan D. Velásquez,
Goldina Ghosh, and Soumya Banerjee

Abstract. Web Opinion Mining (WOM) is a new concept in Web Intelligence. It embraces the problem of extracting, analyzing and aggregating web data about opinions. Studying users' opinions is relevant because through them it is possible to determine how people feel about a product or service and know how it was received by the market. In this chapter, we show an overview about what Opinion Mining is and give some approaches about how to do it. Also, we distinguish and discuss four resources from where opinions can be extracted from, analyzing in each case the main issues that could alter the mining process. One last interesting topic related to WOM and discussed in this chapter is the summarization and visualization of the WOM results. We consider these techniques to be important because they offer a real chance to understand and find a real value for a huge set of heterogeneous opinions collected. Finally, having given enough conceptual background, a practical example is presented using Twitter as a platform for Web Opinion Mining. Results show how an opinion is spread through the network and describes how users influence each other.

5.1 What Is Web Opinion Mining (WOM)?

On many occasions making a good decision requires the opinion of a third person, whether because of insecurity, needing a backup or not having sufficient knowledge

Edison Marrese Taylor · Cristián Rodríguez O. · Juan D. Velásquez
Web Intelligence Consortium Chile Research Centre, Department of
Industrial Engineering School of Engineering and Science, University of Chile,
Av. República 701, Santiago, Chile, P.C. 837-0720
e-mail: {emarrese,crodriguezo}@wi.dii.uchile.cl,
jvelasqu@dii.uchile.cl

Goldina Ghosh · Soumya Banerjee
Department of Computer Science,
Birla Institute of Technology,
Mesra, India
e-mail: goldinag@gmail.com,dr.soumya@ieee.org

J.D. Velásquez et al. (Eds.): Advanced Techniques in Web Intelligence-2, SCI 452, pp. 105–126.
springerlink.com

of the subject. One then begins to consult for information, details, comparisons and opinions in order to have a better idea on the proposal or concept at hand.

For example, wanting to buy a bike is often consulted on with other people who are more related to or have more experience on the subject, for example regarding which brand is best, what characteristics to be considered, which is more convenient -speed or mountain- and if it is better with or without shocks. After considering all the opinions given in this regard, we eventually make a decision on which bike to buy.

If the foregoing advice is considered in a business plan, it shows that for a customer to be sure about what he is going to consume, either products, services, etc., and avoid spending money needlessly or in error, it is essential to consult someone who has experience in the area. The result is the concrete idea that opinions are one of the most important indicators of personal decisions when purchasing a product, taking a tour, selecting a hotel to stay in, where to eat, etc. Many people ask their friends or family to recommend products based on their previous experiences. But there are actually more ways to communicate between persons, considering how thanks to the spread of the Internet and the continued growth of social networks like Twitter, Facebook, and other sites such as blogs or product review pages, we can now take these opinions and experiences from a bigger circle of people than just family or friends. In fact, more people check the opinions of other shoppers before buying a product, when trying to make a good decision [22] [25] [18].

Indeed, based on a survey of more than 2,000 U.S. Internet users [5], more than 75% of product review users reported that the review had had a significant influence on their purchase. Consumers reported a willingness to pay from 20% to 99% more for a 5-star-rated item than a 4-star-rated item. In another survey of 475 U.S. consumers[19], over 60% utilized on-line opinions when making purchase decision. More than 59% of consumers used the web to read on-line reviews, ratings of products or brands and research products and features, when buying products which cost between less than $100 and more than $1000.

The interest in user feedback about a product or service and the influence it has on them is very important for companies that develop products and services as well you can control how their products and their competitors' products were received by the market. As a result, you can determine what things are important to users, what features should improve, modify its advertising and many other things that can mean attracting more users to your brand.

On the other hand, views on political decisions or choices are also interesting for politicians since it allows them to evaluate how things are going, what the most important problems to be solved for the people are, whether they are likely to be elected, and so on.

For these reasons it is interesting to create a tool that can extract a set of opinions and determine what people think about certain products, services, features or be able to understand what the feelings of the people are for a politician based on the amount of positive or negative views people have on any of these topics. Depending only on the target object that has been evaluated, the term opinion mining appears in a paper by Dave et al. [6] where the ideal opinion mining tool should be to "*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)."

5.2 How to Do WOM?

WOM is a new tool and has a long road ahead. Thus, giving a unique definition for WOM is not a simple task because the process' final objective is still unclear. It is possible to find many ways to view this problem in literature. Document Level Opinion Mining and Aspect-Based Opinion Mining are reviewed in this chapter because we consider these to be the most advanced ways to generalize a structured method to do WOM, even though a lot of other perspectives exist.

5.2.1 Aspect-Based Opinion Mining

According to Bing Liu [13], opinions on the Internet can be expressed about anything, e.g., a product, a service, an individual, an organization, an event, or a topic, by any person or organization. This data from Web pages and social media could be structured text or unstructured text. The challenge is transforming the unstructured text into structured text in order to evaluate the positive, negative or neutral sentiment of opinions under study.

He defines an opinion as a quintuple $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ where e_i is the name of an *entity* denoting the target object that has been evaluated, such as a product, service, person, event, organization or topic, and a_{ij}, is an *aspect* of e_i meaning the components and attributes of the target object. Take for example a laptop which has a set of components, e.g. monitor, battery, CPU, and a set of attributes, e.g. size and weight. The components also have their own attributes, e.g. monitor resolution, processing capability and so on. In this model oo_{ijkl} is the orientation of the opinion about aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The opinion orientation oo_{ijkl} can be positive, negative, or neutral or be expressed with different strength/intensity levels.

The previous definition aims to achieve the transformation of unstructured text to structured text and thereby to perform qualitative and quantitative extraction of each of the views. This quintuple gives us the specific information necessary to create a database which is easier to manage. To effect the quintuple process generation, the following tasks must be performed:

1. Extracting from an unstructured opinion, entities and their synonyms grouped in a single cluster. Each entity expression cluster indicates a unique entity e_i
2. Extracting the aspects associated with each of the previously-extracted entities and the grouping of those aspects in a single cluster. Each aspect expression cluster of entity e_i indicates a unique aspect a_{ij}.

3. Extracting the authors of the opinion (holders) and the time the comment was made.
4. Finding whether the orientation of the opinion is positive, negative or neutral.
5. Creating the quintuple of each review to ensure that the entity, aspect, holder, time and orientation of opinion are well-related based on the tasks previously performed.

This is shown in the following example given by Bing Liu [13] of his approach:

Posted by: bigXyz on Nov-4-2010: (1) I bought a Motorola phone and my girlfriend bought a Nokia phone yesterday.(2) We called each other when we got home.(3) The voice of my Moto phone was unclear, but the camera was good.(4) My girlfriend was quite happy with her phone and its sound quality.(5) I want a phone with good voice quality.(6) So I probably will not keep it.

Task 1 should extract the entity expressions, "Motorola", "Nokia", and "Moto", and group "Motorola"and "Moto"together, as they represent the same entity. Task 2 should extract the aspect expressions "camera", "voice", and "sound", and group "voice"and "sound"together, as they are synonyms representing the same aspect. Task 3 should find the holder of the opinions in sentence (3) to be bigXyz (the blog author) and the holder of the opinions in sentence (4) to be bigXyz's girlfriend. It should also find the time when the blog was posted, which is Nov-4-2010. Task 4 should find that sentence (3) gives a negative opinion of the voice quality of the Motorola phone but a positive opinion of its camera. Sentence (4) gives positive opinions of the Nokia phone as a whole and also its sound quality. Sentence (5) seemingly expresses a positive opinion, but it does not. To generate opinion quintuples for sentence (4), we also need to know what "her phone"is and what "its"refers to. All these are challenging problems. Task 5 should finally generate the following four opinion quintuples:

(Motorola, voice_quality, negative, bigXyz, Nov-4-2010)
(Motorola, camera, positive, bigXyz, Nov-4-2010)
(Nokia, GENERAL, positive, bigXyz's girlfriend, Nov-4-2010)
(Nokia, voice_quality, positive, bigXyz's girlfriend, Nov-4-2010)

What makes the WOM proccess difficult, it is the fact that each of the above tasks has not yet been resolved. And to make matters worse there is some information delivered implicitly by opinions. It is still a challenge to ensure that the quintuple has a correspondence to each of its elements. But thanks to this whole process we are able to summarize the information of hundreds of thousands of opinions and determine what people feel about a product, service, etc.

Aspect-Based Opinion Mining is quite important in the opinion-mining area, since it shows an interesting way of ordering information for later analysis of large amounts of data and the ability of the quintuple to receive any other items necessary when performing a specific study. It is a methodology that somehow generalizes the process of opinion mining.

5.2.2 Document Level Opinion Mining

A different approach is proposed by [7], considering a three-phase process which is used by us to create an example of opinion mining in analyzing the influence of some users over others in the social network Twitter section 5.5.

The first phase is Corpora Acquisition Learning, whose aim is to automatically extract documents containing positive and negative opinions from the Web, for a specific domain. They propose collecting the corpus by running queries in a search engine, entering queries specifying the application domain, a seed word they want to find and the words they want to avoid (denoted by the minus "-"sign) and save a determined number of documents from the query results.

The second phase is Adjective Extraction. In relation to this task, they propose an automatic extraction of sets of relevant positive and negative adjectives. The underlying assumption is that adjectives are representative words for specifying opinions, and to achieve extraction, they apply POS (Part of Speech) tagging to recognize adjectives. They then search for associations between the adjectives contained in the documents and the seed words in the positive and negative seed sets, trying to determine whether any new adjectives are associated with the same opinion polarity as the seed words. After that, a filtering process is applied, to keep only the adjectives that are strongly correlated with the seed words. They retain rules containing more than one seed word and then consider adjectives appearing in both the positive and the negative list, applying a formula to rank adjective associations and then deleting the irrelevant ones at the end of the generated list.

The final phase is Classification of new documents using the sets of adjectives obtained in the previous phase. In order to do that, they calculate the document's positive or negative orientation by computing the difference between the number of positive and negative adjectives encountered, from both of the previously described lists. If the result is positive, the document will be classified as positive (the same is true for negative). Otherwise, the document is considered to be neutral. For this phase, it is worth mentioning a proposed method extension, which considers polarity inversion words, such as "not, neither "and "nor".

5.3 Sources for Web Opinion Mining

Having already defined the problem of Web Opinion Mining, the natural next question is related to the possible applications of this technique and subsequently, the set of possible web data sources to use as input. As the Web continues to grow, the number of possible sources for Web Opinion Mining grows rapidly too. Particularly, in the context of WOM, the explosive development of social media plays an important role. Within social media, it is possible to find a variety of different platforms whose content is being increasingly used by individuals and organizations for their decision making [13].

Social media includes web pages such as reviews, forum discussions, blogs, and social networks, like Facebook, Twitter, Foursquare and so on. This variety of sources suggests a heterogeneous mix of structures to work with. As a consequence of this, one needs to specify a different strategy for each source, each one oriented to the particular problem of extracting data, then processing this data and finally discovering valuable information locally within the selected source. Thus, given the mosaic of different structures and procedures for each source, the problem of designing a global methodology is complex and hasn't been widely discussed in literature. The more complete approach to this topic is the idea presented by Bing Liu, from Illinois University, in some publications and with more detail in the book *Web Data Mining*. A review of his work is presented in section 5.2.1.

The rest of this section is a review of the WOM problem in some particular sources. In the first place, the problem of extracting opinions from blogs, forums and news is studied. Then, Twitter is analyzed in order to characterize its content and determine how it can alter the Web Mining strategy to the extraction of knowledge from opinionated documents. Finally a general and brief characterization of other sources is presented.

5.3.1 Blogs, News and Forums

The sources that are probably the richest in opinionated documents are blogs, news and forums. These sources present some common features that make them a good choice when deciding where to look for opinions, but before any analysis, it is worth mentioning that there are some differences between these sources. News and blogs are two important sources of opinions. The writing of the former is formal in comparison to that of the latter since blog articles expressing personal opinions are often written in casual style. Because of this, generally speaking, news documents are more objective, while blog articles are usually more subjective. Other important differences refer to opinion holders, which probably belong to different social classes. Opinions extracted from news are mostly from well-known people, while opinions expressed in blogs may come from a "no name". This issue turns out to be even more important in forums, where it is often difficult to determine the real opinion holder's name [12].However, when analyzing a specific public issue, listing opinions from different sources in parallel can provide many views of the issue, which helps to understand how different social actors react toward the same situation. On the other hand, the most important common feature between news and blogs refers to the document extension, which compared with forums and other sources is much longer. At first glance, this could be seen as an advantage, but long document extensions present a new main problem in Opinion Mining, i.e., how to determine where an opinion sentence is? To solve this problem, major topic-detection techniques are proposed in [12], [7], and some other related publications, to capture main concepts embedded implicitly in a relevant document set. In relation to forums, document extensions can be quite different among singular topics or communities, so it is

difficult to establish a main tendency within this field. However, a useful and positive feature that appears in a high number of review forums is the post-rating system. Post ratings can be used as a predictor of the content's semantic orientation or to contrast and validate text processing and analysis results.

5.3.2 Twitter

Twitter is often considered a microblogging platform, but it is also frequently included as a social network. Given the features of this platform, probably both of these considerations have a certain degree of truth, but for the problem of Web Opinion Mining, the fact that Twitter is a microblog is highly relevant. Microblogging is a growing popular communication channel on the Internet, where users can write short text entrances in a public or private way (to a group of contacts). Messages are extremely short, allowing users to write a maximum of 140 characters on each post, called a tweet. These tweets can be written through the Twitter web interface or through a variety of mobile devices, like smartphones, some cell phones and other devices. These short messages can be seen as being a newspaper headline and subtitle, which makes them easy to produce (write) and process (read). This feature makes microblogs unique when compared to other similar web communication media, like blogs and web pages [2]. As a microblogging social network, the first relevant feature of Twitter's messages are their brevity, which make users look for various special ways to add content in the messages. The most-used approaches are adding content indirectly or trying to use fewer characters to express the same ideas. One important fact on the first topic is the inclusion of links to other web sites to indirectly complement the content of the tweet. In addition to this, it is possible to find a high density of messages containing short URL services, which in fact were created to help Twitter users to include these links with a low number of characters. Another issue refers to the use of Twitter's special characters, like hashtag (#) to denote a particular topic or to call a user, and RT, short for retweet. On the other hand, a different problem associated with tweets is that they are written in colloquial or informal language. In addition to the brevity of messages, this supposes the use of many colloquial symbols or expressions that, in order to be understood as regular text, require preprocessing. In relation to this issue, it is possible to find four different main features:

1. One first special aspect of tweets is that they often include a variety of emoticons. These emoticons help users to express their ideas, feelings or moods in fewer characters, but have a deep impact on text processing. As emoticons are not considered words, they do not have any structure that helps in extracting their lexical meaning, and are not formally included in any dictionary or language that helps to understand their meaning. Nevertheless, as proposed in [20], emoticons can be successfully used to previously determine the tweet's sentiment orientation, opening a wide new field of investigation. Based on this proposal, in [16] emoticons are used to create a corpus collection strategy, defining

two kinds: happy (including “=)”, “:)”,“:D”and others similar) or sad (“:-(”, “:(”, “=(”, etc.)
2. In the second place, spelling mistakes are also common in tweets. This implies the implementation of a preprocessing task in order to fix possible mistakes and ensure a correct interpretation of the content. In relation to this, [11] proposes a technique based on the Levenshtein algorithm, which determines a notion of distance between the misspelled and actually-meant word. This can be used combined with a dictionary to rank and then select the most probable letter replacements from a list of previously-generated possible word candidates.
3. Another feature commonly found in tweets is the repetition of one letter in a word. This is mostly done when users want to add some emotive intensification to the text, for instance repeating vowel letters as in “I loooove you”. In this field, [1] proposes a control system based on regular expressions for Spanish with back reference, replacing the appearance of two or more characters by only one letter, excepting the groups “cc”, “ll”and “rr”, which are commonly used in this language.
4. A last feature is related to the use of Internet language, or “Netspeak”, in the messages. In this context, the use of capital letters can be problematic when tokenizing or lemmatizing text during preprocessing.

Finally, it is interesting to see Twitter as a network of related users, who share opinions among themselves and influence each other. Twitter provides some useful tools that can be used to analyze how a specific topic or opinion tendency is spread through the network, and discover users who influence others or are more easily influenced by another. Based on this, one could be able to, for instance, define clusters and find non-trivial segmentations based on the characterization. A proposal on this field is presented in section 5.5.

5.3.3 Other Media

There are a lot of other possible sources for WOM. In the context of social networks, Facebook is often proposed as a powerful and complete repository. Nevertheless, the main problem in this case is related to privacy policies and access limitations to the contents. Thus, opinionated documents are not always publicly available, and it is difficult to reach them.

5.4 Visualization Techniques

In most of the cases, studies need to analyze a large number of opinion holders. Common sense indicates that one opinion from a single holder is usually not sufficient for action. This idea naturally leads to the task of opinion summarization. The literature proposes some different approaches to summarizing and then

visualizing summarized opinions. This section focuses on different visualization techniques which require, in one way or another, some kind of previous summarization. This last task is a complex and pretty well-studied field. Its application to opinions (and web opinions) is just a particular case and will be briefly discussed in this section.

As presented in [3], traditional summarization consists of constructing new sentences from the opinionated document, in order to extract the document main points. In [3], the opinion summarization technique proposed is founded on the idea of analyzing relationships among basic opinionated units within the document. More precisely, the paper presents an approach to multi-perspective question answering (MPQA) that views the task as one of opinion-oriented information extraction. Briefly, the information extraction system takes as input an unrestricted text and summarizes the text with respect to a previously-specified topic or domain of interest, finding useful information about the domain and encoding that information in a structured form, suitable for populating databases. The process involves creating low-level annotations of the text which are then used to build the summary. Visualization in this context is thus the construction and presentation of short sentences (or sets of sentences) that capture a document's main opinionated ideas.

The traditional fashion for summarization then means producing a short text summary that gives the reader a quick overview of what people think about a defined object. Some traditional summarization techniques can be found in [4], [15], [21] and [23]. Nevertheless, the main weakness of these text-based summaries is that they are just qualitative, which means that it is not possible to apply any numerical or quantitative analysis to them. As proposed in [13], the quantitative side is crucial, just as in traditional survey research.

In this context, opinion quintuples defined by Liu's approach are a good source of information for generating both qualitative and quantitative summaries, and can be stored in database tables. Based on this, a kind of summary based on aspects is defined, which is called aspect-based opinion summary [9], [10]. Having built the proposed structure, a whole set of visualization tools can be applied to see the results in all kinds of ways, to then gain insights into the opinions. In this case, bar charts or pie charts are both used. As an example, data can be visualized using a bar chart in which each bar above the X-axis shows the number of positive opinions on one aspect, and the corresponding bar below the X-axis shows the number of negative opinions on the same aspect. A different technique may only consider showing the percent of positive opinions.

Liu's visualization proposal is also interesting because it enables comparison of opinion summaries of some competing products. In addition to this, instead of just generating a quantitative summarization, a text summary directly from input reviews is also possible, generating natural language sentences based on what is shown in the charts [14].

It is important to mention that this technique is only related to product opinions and results quite differently from traditional text summarization because it focuses and mines only the features of these products, while also determining whether the opinions are positive or negative. There is no rewriting of original sentences to

capture the main points of the opinionated selected document, as in the classic text summarization described previously.

On the other hand, a completely different approach in relation to text summarization and visualization is presented in [12]. As presented before, traditional summarization algorithms rely on the important facts of opinionated documents and remove redundant information. Nevertheless, it is likely that sentiment degree and correlated events play major roles in summarization. Because of that, [12] proposes that repeated opinions of the same polarity cannot be dropped because they strengthen the sentiment degree, but repeated reasons why they stated a position should be removed when generating summarization. To apply this summarization system it was therefore needed to know which sentences were opinionated and then decide if they focused on a designated topic. An algorithm that detects and extracts major topics in long documents and then classifies them in positive or negative orientation in relation to that topic was then developed. Then, for brief summarization, the document with the largest number of positive or negative sentences is picked up and its headline is used to represent the overall tendency of positive-topical or negative-topical sentences. For detailed summarization, a list of positive-topical and negative-topical sentences with higher sentiment degree is generated.

As a complement, an opinion-tracking system that shows how opinions change over time is proposed. The tracking system is very similar to Liu's proposal and consists of bar charts that simply count the number of positive-topical and negative-topical sentences on a selected topic at different time intervals. Nevertheless, a large number of relevant articles is required.

Finally, Tateishi's approach is worth mentioning, which introduces radar charts for summarizing opinions and has been frequently cited in the literature. A more detailed description of this technique can be found in [24]. Sadly, as the original document is only available in Japanese as this text is being written, it was not possible to include a deeper analysis due to language issues.

5.5 An Application of WOM in Twitter

This example envisages the social network Twitter. The record set provides us with some vital information such as user ID, number of followers, number of following users, number of tweets and frequency of tweets. There can be both important and unimportant information that can be shared. This information can be transmitted through the blogs where some senders take up the initiative to transform different kinds of events. Based on the blogs the person can have followers. Followers are those who either agree with an opinion or get influenced by other's events and propagate these messages to yet others. Again, if we consider the number of blogs or tweets and their values, then the number of influential members of the tweeter social network can also be identified. Based on the tweeter database it is possible to cluster the number of followers and following individuals. By applying the

subtractive clustering method in fuzzy logic the grouping of those members who are more influential is done.

5.5.1 Extracting Significant Information from Tweets

In this section a methodology for extracting and processing tweets is proposed by using *Subtractive Clustering using fuzzy logic* with the application of logical operations and the *if-then rule* statement. The relevant terminologies are as follows:

- **Fuzzy Inference System:** This is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping process can then be applied to make decisions or perform any pattern analysis. The Fuzzy Inference System also involves the concepts of membership, logical operations and the if-then rule.
- **Membership functions:** Membership functions are curves that define how each point in the input space is mapped to the degree of membership between 0 and 1. The input space is also called 'Universe of Discourse'.
- ***Logical operations*** like AND, OR, NOT are also applied.
- 'If-then rule'statements are used for conditional statements that are comprised of fuzzy logic.
- ***Subtractive Clustering:*** This method of clustering functions by assuming each data point to be a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center based on the density of the surrounding data points. Subtractive clustering is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. The cluster estimates obtained can be used to initialize iterative optimization-based clustering methods and model-identification methods. The steps in performing the process are as follows:
  - Selecting the data point with the highest potential to be the first cluster center.
  - Removing all data points in the vicinity of the first cluster, called a *radii*, in order to determine the next data cluster and its center location
  - Repeating this process until all of the data are within the *radii* of a cluster center.

The *radii* are a vector of entries between 0 and 1 that specify a cluster center's range of influence in each of the data dimensions. The best value of a given *radii* is usually between 0.2 and 0.5.

A *Sugeno-type fuzzy inference* system that models the data behavior is generated to provide a fast, one-pass method to take input-output training data. A typical rule in a *Sugeno fuzzy* model has the form,

'If Input 1 = x and Input 2 = y, then Output is $z = ax + by + c$'.

For a zero-order Sugeno model, the output level z is a constant ($a = b = 0$). The output level z_i of each rule is weighted by the firing strength w_i of the rule. The final output of the system is the weighted average of all rule outputs, computed as:

$$FinalOut = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \tag{5.1}$$

where N is the number of rules.

5.5.2 Data Quality Evaluation

The first step in any information extraction process is the evaluation of the data quality to be processed. In the case of text originated in Twitter, these can content many wrong words, typos mistakes, emoticons[1] , non structured sentences, etc. A second step is to select the algorithms to use for processing the Twitter data and finally, a mathematical treatments of the extracted data for detecting the most influential members of the Twitter community.

5.5.2.1 Components of Standard Data Set

To validate the proposed analytical solution, we collect the data from the Twitter social networking site. The database is extracted using the three main streaming products: The Streaming API, User Streams and Site Streams.

Streaming API: It is used to collect public statuses from all users, filtered in various ways. It can be by user ID, by keyword, by random sampling, by geographic location and other parameters.

User Streams: All the data are required to be updated. It provides public and protected statuses from followings, direct messages, mentions, and other events taken on and by the user.

Site Streams: This allows multiplexing of multiple User Streams over a Site Stream connection.

While using the Search API one is not restricted by a certain number of API requests per hour, but instead by the complexity and frequency. As requests to the Search API are anonymous, the rate limit is measured against the requesting client IP. The record set collected contains certain information about 99 unique members of the Twitter community where each member's 'number of followers', 'number of following', 'number of tweets'and 'frequency of tweets'are present. Table 5.1 shows the details of the record set:

[1] An emotion can be defined as "feeling states with physiological, cognitive, and behavioral components" [8], in that sense an emoticon can be defined as a short sequence of keyboard letters and symbols for expressing a human emotion in chats, blogs, e-mails, twitter, etc.

Table 5.1 The detailed record set of Twitter

user_id	no_of_followers	no_following	no_of_tweets	tweet_frequency
156518278	66	80	4248	48.116505
33765767	682	660	11106	74.75192
450587971	32	65	297	36.473682
380242792	7	0	445	20.493422
20531842	279	392	13876	87.58521
55989455	611	337	51513	376.39978
83018127	0	0	38	0.3089431
224300080	61	1	55886	883.0745
27937504	1285	465	12547	82.857544
467461925	31	0	1331	251.81082
467464356	21	0	1112	210.37839
163777200	587	439	23128	270.7291
402424165	2	0	298	17.98276
180749032	141	145	9097	114.94404
18370866	465	190	6554	39.65255
14115513	125	25	547	2.6479945
144538924	1034	244	14456	155.91988
20155063	165	117	1519	9.544884
203226606	100	0	26745	376.69016
102953286	191	195	6766	60.876606
132729891	2805	117	1265	12.9838705
214006837	458	373	20129	298.5233
221106075	32	71	4446	68.85398
224996941	127	109	5610	88.84615
40540088	170	286	1916	13.226824
330240789	120	204	1453	43.465813
465856538	3	9	16	2.8717947
430139433	2	14	178	15.575
93816436	624	396	26756	229.52452
467452745	32	0	961	181.81082
281369631	32	42	228	5.018868
76161854	0	0	49	0.38713318
124483151	7	20	25	0.24752475
401097105	1	0	29	1.720339
234992003	1043	864	42475	718.17633
80854705	902	1936	782	6.2991943
348272136	957	1549	146	4.985366
138510278	124	159	2909	30.575073
274335070	6	7	720	15.180723
108578780	386	383	1249	11.503947
289726223	522	10	18871	437.4073
213000229	631	0	10536	155.59494
456494008	28	1	1409	197.26
468635975	0	11	4	0.7777778
213000229	631	0	10536	155.59494

5.5.2.2 Processing Data Originated in Twitter

To design the data processing stage, the following steps need to be accomplished:

A Comparison among the 'number of followers'and 'number of following'data of each member to find out which ones all are influential and which ones all are influenced. The condition is that if number of followers > number of following, then a member is categorized as being influential and if not, then as being influenced.
B Based on the number of tweets the tweet weight age is judged by providing a threshold value and then analyzing whether the numbers of followers are directly proportional to the number of tweets made by each member.
C If the numbers of tweets along with their weighted value are high, then those members can have larger number of followers. This measurement is done by checking the direct proportionality among the number of tweets and number of followers. If it is inversely proportional then the number of tweets was not influential.
D An analysis report that is derived from the above two conditions will help in clustering those members who are more influential than the others based on the data set of Twitter in Table 5.1 by using the subtractive clustering technique in fuzzy logic.

5.5.2.3 Modeling the Twitter User Behavior

Certain information is available from the Twitter database, including each individual having an ID number, along with their number of followers and following. Based on these two fields, an analytical report of the person's nature can be identified to determine whether the person is influential or influenced. Based on these data a very simple analysis can be derived that is a symmetric relation. Symmetric relation can be expressed as "*if x is related by R to y, then y is related by R to x*".

Here, the numbers of individuals (unique members of the Twitter community) are denoted as,

$$I = \{I_1, I_2, I_3, ..., I_n\}$$

and the nature of the individual can be categorized in the following form,

$$Nature = \{Influencial, Influenced\}$$

By applying the concept of symmetric relation the inference that can be drawn is: If number of followers > number of following, then member is categorized to be influential and if not then influenced.

Analyzing fields which provide information about the number of tweets made by each member can help in judging the rate of increase or decrease in followers. The number of followers will be directly proportional to the number of tweets sent, if the tweets are larger in number and also the weightage of each tweet is high, which

means the message is important and valuable. This property can be represented in the following format:

$$\begin{aligned} NF_i &= NT_j * k, \\ i &= \{i_1, i_2, ..., i_n\} \\ j &= \{j_1, j_2, ..., j_m\} \end{aligned} \tag{5.2}$$

where NF= Number of Followers, NT= Number of Tweets, k= Constant.

The reverse representation happens when the numbers of tweets are less in number and when they are of less importance. This means that the number of followers will be inversely proportional to the number of tweets. This property is represented in the following way:

$$\begin{aligned} NF_i &= \frac{k}{NT_j}, \\ i &= \{i_1, i_2, ..., i_n\} \\ j &= \{j_1, j_2, ..., j_m\} \end{aligned} \tag{5.3}$$

Where, NF= Number of Followers, NT= Number of Tweets, k= Constant

Thus from equations 5.2 and 5.3, we can detect the most influential members of the tweeter community based on the database of the particular moment by applying subtractive clustering. The member with the highest potential is selected to detect the first cluster, and then all the points are removed from the vicinity of the first cluster so as to detect the next cluster. This process is repeated again and again so as to make a collection of all the members that are influential as well as whether their tweets are important or of high value. The data behavior that can thus be identified from the record set is a one-pass method to take input-output training data and generate a Sugeno-type fuzzy inference system.

5.5.3 Analysis and Results

The effectiveness of measuring the different levels of followers is determined in Figures 5.1 and 5.2. In Figure 5.1 the different rankings of members are plotted based on the number of followers less than the number of following. The x-axis denotes the different number of followers and y-axis denotes the different ranking of the followers based on the different tweet weight age.

In figure 5.2 the different rankings of members are plotted based on the number of followers greater than following.

Depending on the number of tweets and also the valuation of those tweets the number of followers is denoted. Figure 5.3 and figure 5.4 show the scattered plotting of those members whose tweets were important and collected a larger number of followers whereas the members who had less important tweets had a lesser number

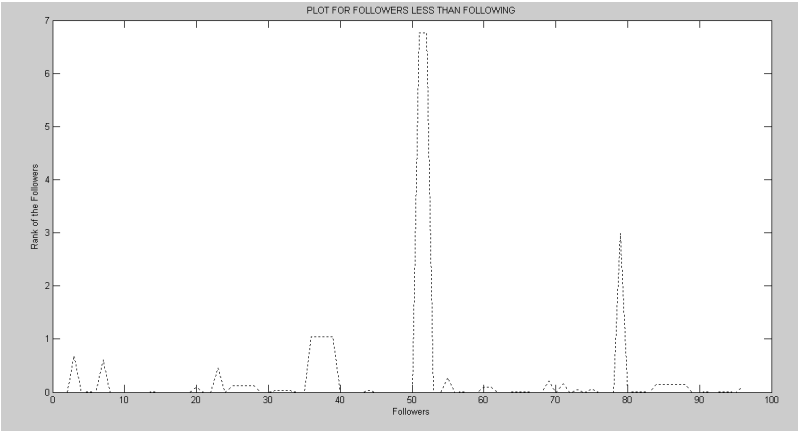

Fig. 5.1 Members with followers less than following

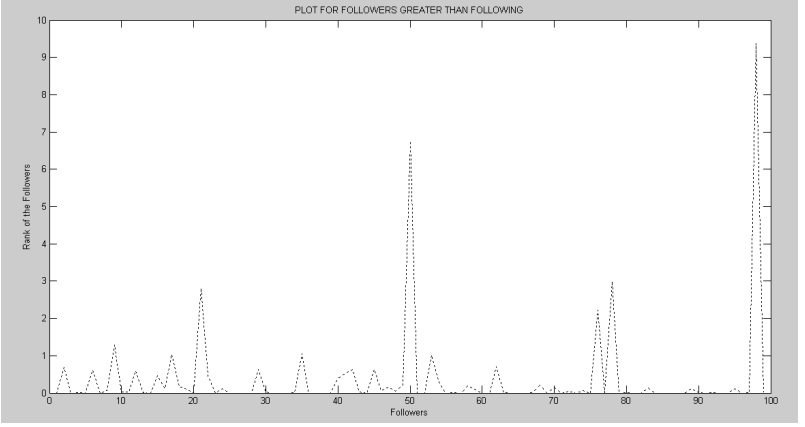

Fig. 5.2 Members with followers greater than following

of followers. Figure 5.3 describes the collection of those members who had higher numbers of followers due to their strength in tweets.

Similarly, figure 5.4 denotes the reverse condition where the numbers of followers are less due to the less important tweets.

Hence, based on these results, we can draw a conclusion about those members who are very influential and shared a lead role. The following result is plotted in figure 5.5 by applying the subtractive clustering technique.

The social network is a place that helps in sharing different opinions with different users. In this way plenty of information could be provided about knowledge being shared. Additionally, in those cases where some vital tweets are made by some individual then those tweets are given much more importance than the

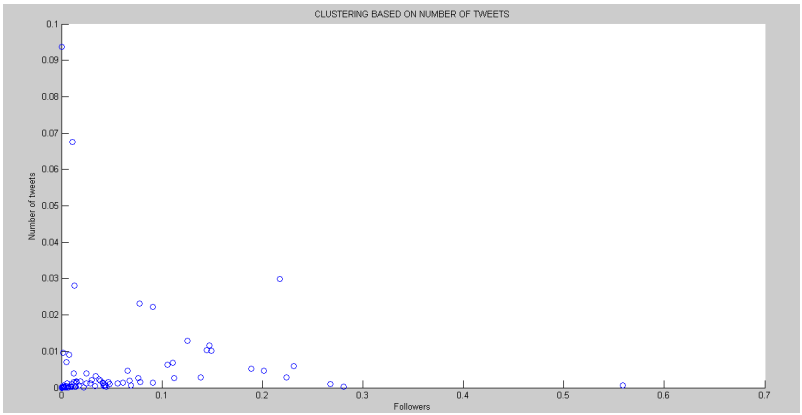

Fig. 5.3 Important tweets with influenced followers

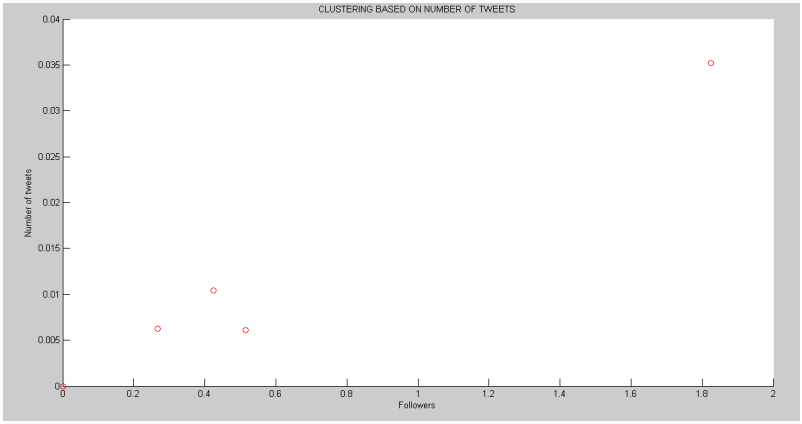

Fig. 5.4 Unimportant tweets with less influenced followers

others, since they get propagated among more followers. This can be represented in a graphical representation where the x-axis denotes the number of individuals transferring tweets and the y-axis represents the flow of the tweets. This scattered graph denotes when the followers rate was high due to the high weight age of the tweet. The number of followers rate increases based on time, event and the members of the tweeter group. The rate of increase in flow of the tweets among the individuals can also be expressed mathematically, $y = x^a$, where x = weight of the tweet (high-valued tweets will be transmitted to many individuals) y = number of followers following that tweet and transferring them to others (depending on the tweet value) a = depends on the previous value of y.

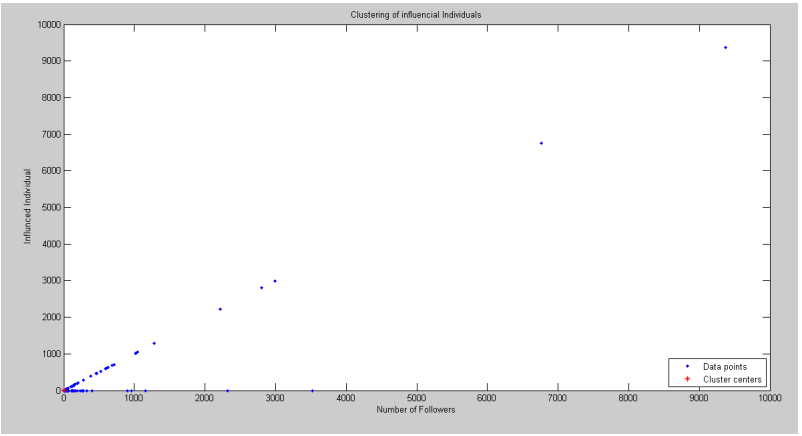

Fig. 5.5 Subtractive clustering of all influential members

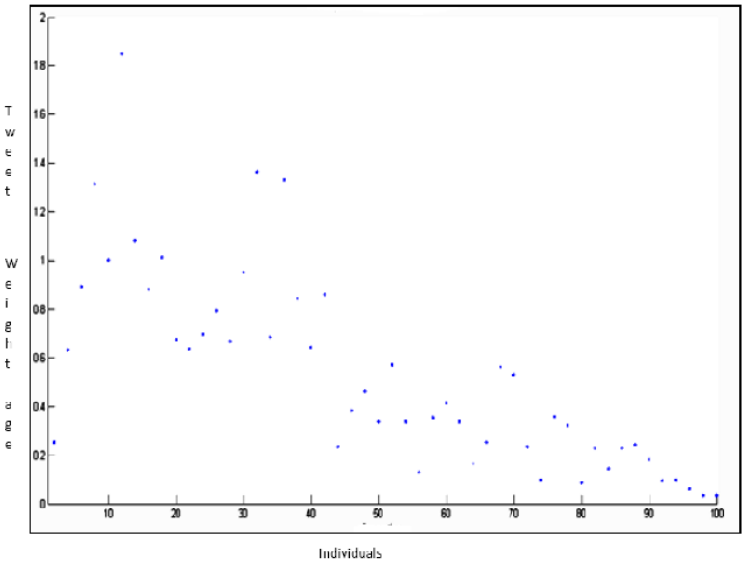

Fig. 5.6 Scattered plot denoting flow of influential tweets

The most influential person of a group can always be ranked as the best informer of some important event, news or for sharing good thoughts. In Figure 5.7 bar diagram denotes here the best member, who had communicated well and had been placed in the high rank. This is followed by the ranks of the other members' as per their weight of the tweet. The x-axis denotes the individual members and the y-axis represents the rank level of each individual where they belong.

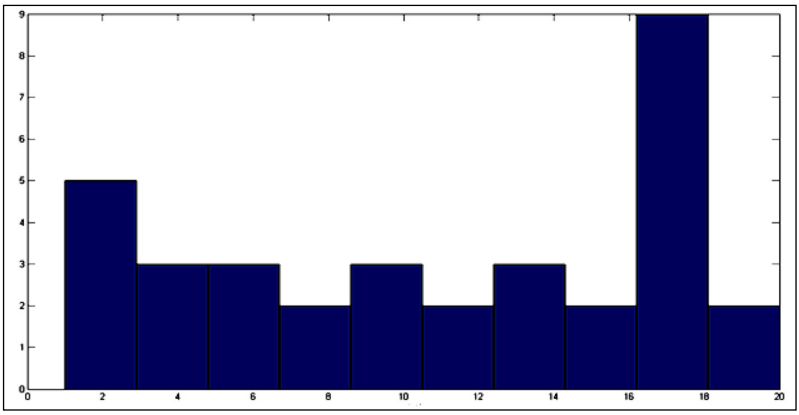

Fig. 5.7 Histogram denoting rate of flow of influential tweets

5.5.4 Tweeter as Novel Web User Behavior Data Source

In fact, the data originated in social networks are new sources for understanding the web user behavior. A very important issue in this kind of data is these normally are generated by own web user decision, corresponding usually what the web user is thinking and feeling, However, these opinion can be influence by other web user opinions, then to analyze what influenceable a social network can became will be always necessary for analyzing the web user behavior.

A social network community like Twitter, the followers receive the others web user influence by tweets an retweets opinions about some particular theme. Then this kind of social site is not only a medium of sharing important events but also in taking part in knowledge sharing and also participating in some conversations. The blogs that are posted by the members of such social network community are some time a mind opener for other members also. This leads in judging whether the blogs posted by the person is valuable which further leads in categorizing the influenced and influencing persons. This can be finally subjected towards many real life applications as sentiment analysis and emotion modeling pertaining to social network analysis.

5.6 Summary

With the inception of Web 2.0 and the explosive growth of social media on the Web, users have now the possibility to express personal opinions about products, services, and therefore, access a huge repository of these opinions to make better decisions about buying or using a product or service. This information could also be important

for enterprises, that might be interested in knowing how their products or services are perceived by the market.

Among the different sources where users add their personal opinions, we distinguished and discussed four: Blogs, News, Forums and the social network Twitter. Each one presents a specific operating mode which brings particular issues related. In section 5.3 we analyzed these four sources and tried to determine how their particular problems can affect the mining process.

Having given a look at all these issues, WOM begins to become a complex and challenging task. Then we stated that developing a system that embraces all these problems is naturally difficult. Nevertheless, it is possible to find various approaches in literature. In section 5.2 we introduce two of these methodologies: Document Level Opinion Mining and Aspect-Based Opinion Mining, even though a lot of other perspectives exist.

Each approach proposes a series of logical steps to transform the complex and heterogeneous data from different sources into a structured and simpler configuration. Nevertheless, giving a unique definition for WOM is not a simple task because the process final objective is still unclear.

On the other hand, in section 5.4 we presented summarization and visualization techniques that allow a better comprehension of the mining process results. For this reason these are remarkably important tasks and deserve to be studied. Multiple techniques involving classic text summaries, chart visualization methods and other new proposals were introduced.

Finally, in order to achieve a better understanding of one Web Opinion Mining existing technique, we presented a practical example in section 5.5, which implements Document Level Opinion Mining in the social network Twitter. In this case, the goal was detecting the most influential users on Twitter community.

References

1. Balbachan, F., Dell'Era, D.: Automatic sentiment analysis of short texts in twitter platform (2011) (in Spanish)
2. Bifet, A., Frank, E.: Sentiment Knowledge Discovery in Twitter Streaming Data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
3. Cardie, C., Wiebe, J., Wilson, T., Litman, D.: Combining low-level and summary representations of opinions for multi-perspective question answering. In: Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, pp. 20–27 (2003)
4. Carenini, G., Ng, R., Pauls, A.: Multi-document summarization of evaluative text. In: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pp. 305–312 (2006)
5. comScore/the Kelsey group. Online consumer-generated reviews have significant impact on offline purchase behavior. Press Release (November 2007), `http://www.comscore.com/press/release.asp?press=1928`

6. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web, pp. 519–528. ACM (2003)
7. Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., Poncelet, P.: Web opinion mining: How to extract opinions from blogs? In: Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, pp. 211–217. ACM (2008)
8. Hsee, C., Hatfield, E., Carlson, J.G., Chemtob, C.: The effect of power on susceptibility to emotional contagion. Cognition and Emotion 4, 327–340 (2004)
9. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
10. Hu, M., Liu, B.: Opinion extraction and summarization on the web. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21, p. 1621. AAAI Press, Menlo Park (2006)
11. Jurafsky, D., Martin, J.H., Kehler, A., Vander Linden, K., Ward, N.: Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, vol. 2. Prentice-Hall, New Jersey (2000)
12. Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, vol. 2001 (2006)
13. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Springer (2011)
14. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the 14th International Conference on World Wide Web, pp. 342–351. ACM (2005)
15. Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G.: Optimizing informativeness and readability for sentiment summarization. In: Proceedings of the ACL 2010 Conference Short Papers, pp. 325–330. Association for Computational Linguistics (2010)
16. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC (2010)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2(1-2), 1–135 (2008)
18. Park, D.H., Kim, S.: The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. Electronic Commerce Research and Applications 7(4), 399–410 (2009)
19. Razorfish. Digital consumer behavior study (2007)
20. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop, pp. 43–48. Association for Computational Linguistics (2005)
21. Seki, Y., Eguchi, K., Kando, N., Aono, M.: Opinion-focused summarization and its analysis at duc 2006. In: Proceedings of the Document Understanding Conference (DUC), pp. 122–130 (2006)
22. Shin, H.S., Hanssens, D.M., Kim, K.I., Gajula, B.: Impact of positive vs. negative e-sentiment on daily market value of high-tech products (2011)

23. Stoyanov, V., Cardie, C.: Partially supervised coreference resolution for opinion summarization through structured rule learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 336–344. Association for Computational Linguistics (2006)
24. Tateishi, K., Fukushima, T., Kobayashi, N., Takahashi, T., Fujita, A., Inui, K., Matsumoto, Y.: Web opinion extraction and summarization based on viewpoints of products (in japanese). Information Processing Society of Japan SIGNL Note 93, 1–8 (2004) (in Japanese)
25. Zhu, F., Zhang, X.: Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. Journal of Marketing 74(2), 133–148 (2010)