Opinion Mining: Issues and Challenges (A survey)

Bakhtawar Seerat
Department of Computer Engineering
National University of Sciences and Technology,
Islamabad, Pakistan

Farouque Azam
Department of Computer Engineering
National University of Sciences and Technology,
Islamabad, Pakistan

ABSTRACT
Opinion mining is crucial for both individuals and companies. Individuals may want to see the opinion of other customers about a product to analyze it before buying it. Companies want to analyze the feedback of customers about their products to make future decisions. So, analyzing customer's opinion and their response is important. Mining is used on product reviews that are available on different blogs, web forums, and product review sites to evaluate opinions of customers. By doing so, new customers are able to find views of others about a product and can decide which product to buy by the help of opinion of customers already using the product. In addition comparison of same feature of products by different vendors is done. In this way companies can focus on improving the features of their product that are not popular among customers. This leads to overcome the requirements of marketing intelligence and product benchmarking in the production industry. In this paper we do a survey of papers and will summarize the issues and challenges of opinion mining that affect the results of opinion mining.

Keywords
Knowledge discovery, Data Mining, Web Mining, Opinion Mining, Sentiment Analysis, Issues, Challenges.

1. INTRODUCTION
It's an emphasis of the marketing places to identify and subsequently satisfy consumer needs. So, in order to examine consumer needs and to implement effective marketing strategies aimed at satisfying these needs, marketing managers need relevant, current information about consumers, competitors and other forces in the marketplace. There is not much study on opinions in the past, as a significant part of consumer information, which is present in company's own databases, has been ignored in the past and there was less opinionated text available before the existence of World Wide Web.

Nowadays users of web 2.0 contribute content actively in web-forums and product review websites. With the growth of the web over the last decade, opinions can now be found almost everywhere-blogs, social networking sites like Facebook and Twitter, news portals, e-commerce sites, etc. Therefore, to get additional market insights, modern companies have a strong need to utilize this user-generated content. So, if a person wants to purchase a certain product or companies want to know opinions of the consumers about their products, web reviews can be used for this purpose as the web has acquired immense value as an actively evolving repository of knowledge for market research. [4, 5]

Due to availability of large volume of information on Web, opinion mining can still be a formidable task. Consider a user looking to buy a laptop. Figure 1 shows all the available laptop reviews for a Dell laptop obtained from Google Product Search. Although these opinions are meant for just one product, there are more than 400 reviews for this one product from around 20 different sources. Such overwhelming amounts of information make summarization of the web very critical. Google searches facts, not opinion. 'Opinion' mainly includes opinionated text data such as blog/review articles, and associated numerical data like aspect rating is also included. To generate a concise and digestible summary of a large number of opinions is the study of Opinion Summarization. The simplest form of an opinion summary is the result of sentiment prediction /opinion mining. [13].

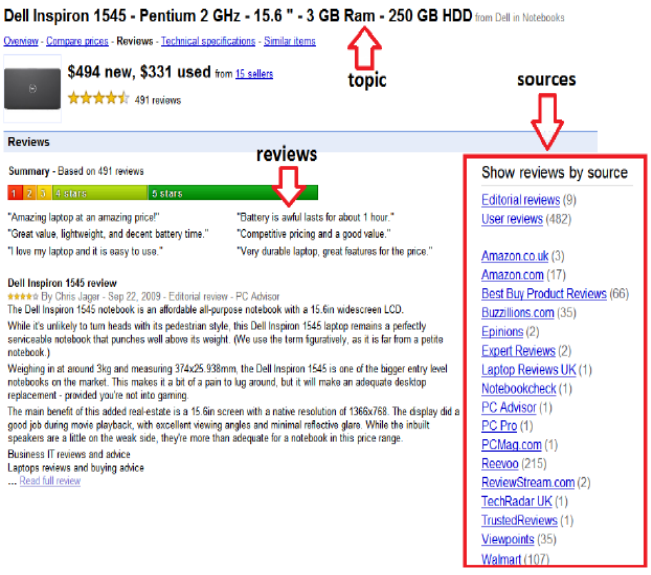

Figure 1: Example Google product search on Dell laptop

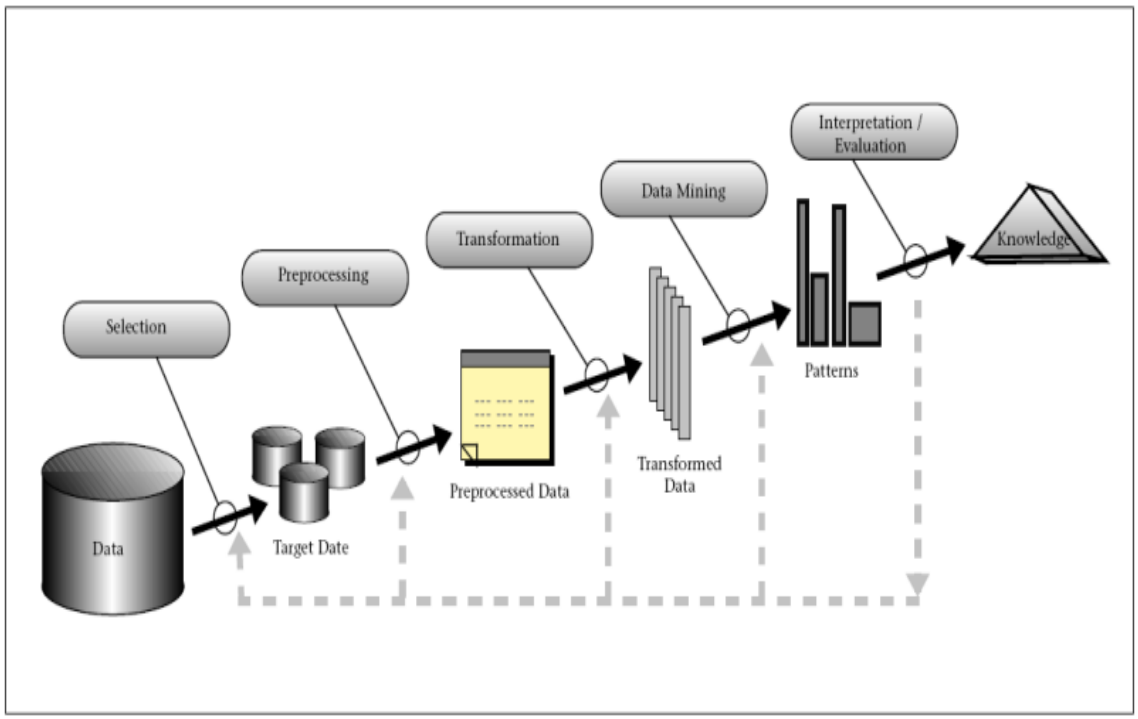

Figure 2: Overview of the steps consulting the KDD process

Opinion mining involves Text Mining and Language Processing (NLP) and Text Classification. Text Mining has a great potential to overcome the current deficiencies. Unfortunately, natural language processing (NLP) encounters a range of difficulties due to the sophisticated nature of human language. Moreover, the area of opinion mining involves the problem of Text Classification, which is totally different from the usual Text Mining. In usual Text Classification, the focus is on identifying topic, whereas for opinion mining, Sentiment Classification is done which focuses on the assessing writer's sentiment toward the topic. Emotions are not satisfactorily analyzed with keyword based methods. [4, 5]

The paper discusses the challenges for opinion mining which include some common issues of KDD process. The techniques used for mining also affect the results. So, they are also an issue affecting opinion mining. Opinion mining has some challenges of its own nature.

KDD process is discussed as follows.

1.1 Knowledge Discovery

In modern database technology, processing of very huge volume of data is involved, to extract new knowledge.

- **Definition(Knowledge discovery)** is the extraction of hidden, previously unknown information from data that is of interest to the end user's knowledge and objectives.[1]

As knowledge is not always represented explicitly, so to extract useful information from complex and large volume of data, current database technology involves processing of data. So, Data mining and knowledge discovery (KDD) process is used for this purpose and the field of data mining and knowledge discovery from databases has emerged as a new discipline in engineering.

1.1.1 Knowledge discovery process in databases (KDD)

The KDD process is outlined in Figure 2. [2, 12]

The KDD process includes the following steps:

1. **Learning the application domain**: includes relevant prior knowledge and the goals of the application.

2. **Creating a target dataset**: includes selecting a dataset or focusing on a subset of variables or data samples on which discovery is to be performed.

3. **Data cleaning and preprocessing**: includes basic operations such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues such as data types, schema, and mapping of missing and unknown values.

4. **Data reduction and projection**: includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. **Choosing the function of data mining**: includes deciding the purpose of the model derived by the data mining algorithm e.g. this can be summarization, classification, regression, and clustering.

6. **Choosing the data mining algorithm(s):** includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over real) and matching a particular data mining method with the overall criteria of the KDD process (e.g., the user may be more interested in understanding the model than in its predictive capabilities).

7. **Data mining**: includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis.

8. **Interpretation**: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.

9. **Using discovered knowledge**: includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge .

In paper [1] some common challenges of KDD and their proposed solutions are presented. These also affect opinion mining.

1.1.2 Issues that affect the overall knowledge discovery process

Main issue in any discovery system is the problems in ensuring the quality (consistency, accuracy, and completeness) of the discovered knowledge. The problems (i.e., incorrect, inconsistent, and incomplete rules) do exist in the discovered rules due to:

- An **inadequate database design:** Knowledge discovery depends upon how well the database is created and maintained.
- **Poor data:** Efficiency of the discovery process and the quality of the discovered knowledge are strongly dependent on the quality of data.
- The **vulnerability/limitations of the tools** used for discovery
- **Flaws in the discovery process**: The process used to obtain and validate the rules using a given tool on a given database.
- **Real-world databases** present difficulties as they tend to be dynamic, incomplete, redundant, inaccurate, and very large. So, these problems associated with the discovery techniques /schemes cause the discovered knowledge to be incorrect, inconsistent, incomplete, and uninteresting.
- **Relational databases** create new types of problems for knowledge discovery since they are normalized to avoid redundancies and update anomalies, which make them unsuitable for knowledge discovery.
- **Operational relational databases**, built for online transaction processing, are generally regarded as unsuitable for rule discovery since they are designed for maximizing transaction capacity.
- **Summary and historical data absent**: Summary and historical data, which are essential for accurate and complete knowledge discovery, are generally absent in the operational databases. Rule discovery based on just the detailed (most recent) data is neither accurate nor complete.

1.1.3 Solution

Some solutions are proposed to avoid these problems.

- **Cleaning data:** Clean data should be provided to the discovery process in order to discover useful information from the databases. Hence, the databases need to be cleaned before the actual discovery process takes place in order to avoid discovering incomplete, inaccurate, redundant, inconsistent, and uninteresting knowledge. **Improved tools and techniques** :Different tools and techniques have been developed to improve the quality of the databases in recent years, leading to a better discovery environment.
- **Data warehouse instead of operational, relational databases:** Most of the knowledge discovery has been done on operational relational. A **data warehouse** is a better environment for rule discovery since it checks for the quality of data more rigorously than the operational data-base. It also includes the integrated, summarized, historical, and metadata which complement the detailed data.
- **Summarized and historical data:** Summarized data contains patterns that can be discovered. Historical data (i.e., sales product 1982-1991) is essential in understanding the true nature of the patterns representing the data. [2]

1.2 Data Mining

Data mining is the analysis step of the KDD process and the overall process is dependent on it. It aims at extracting knowledge from large amount of data in an understandable structure that is useful for companies and individuals. Sifting through very large amounts of data for useful information, Data mining uses intelligence techniques, neural networks, and advanced statistical tools(such as cluster analysis) to reveal trends, patterns, and relationships, which might otherwise have remained undetected. In contrast to an expert system (which draws inferences from the given data on the basis of a given set of rules) data mining attempts to discover hidden rules underlying the data. This is also called data surfing.

1.2.1 Soft computing methodologies for data mining

Soft computing methodologies are used for data mining. These vary depending upon whether the data is structured/semi-structured /non-structured. Some are:

For structured data:

- A decision tree can describe a rule set in the format of a tree structure. The tree is regarded as the set of IF-THEN rules.
- Neural networks and rough sets are employed for rule extraction from data. A neural network can describe a rule set in the format of a network structure. The network stores the relationships between attributes and classes as weights of the arcs in the network. The weights are appropriately adjusted by the back propagation algorithm.
- A genetic algorithm inspired by the concept of evolution can acquire a rule set from structured data. Genetic algorithms are involved in various optimization and search processes, like query optimization and template selection.

For non-structured data:

- Fuzzy sets naturally deal with uncertainty. So, fuzzy set theory comes to deal with ambiguous data**.** [1]

These techniques are a factor that also affects quality of the mined knowledge.

2. WEB MINING

As large volume of information is available online, World Wide Web is a natural area for data mining. The Web mining research is at the crossroads of research from several research communities such as database, information retrieval, and Artificial Intelligence. [1]

As web knowledge is scattered and due to lack of any uniform format, web mining is a difficult task and involves many issues. As in paper [6] and [7], web mining and its categories are explained:

- **DEFINITION (Web mining)** is the process of applying data mining techniques for discovering patterns from the Web.

Web mining is divided into three different types, which are Web usage mining, Web content mining and Web structure mining. These categories are shown in figure 3.

2.1 Web Usage Mining

Web usage mining is the process of discovering what users want to see on the Internet. Some users are interested in textual data while others in multimedia data. This is done by making use of user logs.

2.2 Web Structure Mining

Web structure mining is the process of extracting knowledge from web pages by focusing the structure. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

2.3 Web Content Mining

Web content mining aims to extract useful information from contents of the web page. It involves scanning of all the contents on a web page to find its relevance with the search query [3]

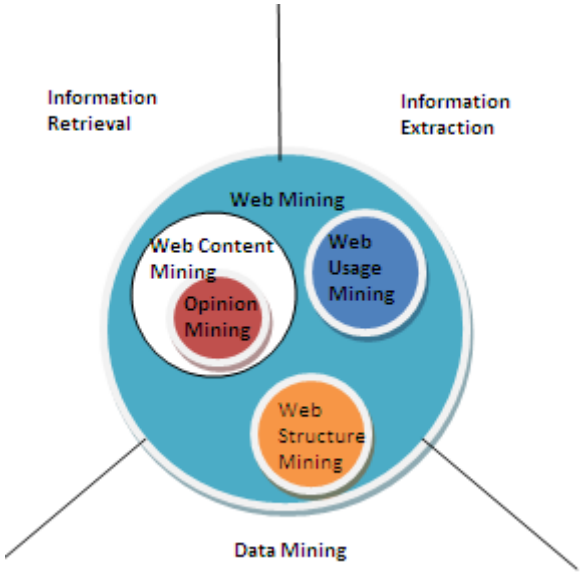

Figure 3: Web mining categories

3. OPINION MINING

It is kind of web content mining. Figure 3 shows this categorization clearly.

- **DEFINITION** If a set of text documents (T) are given, that have opinions on an object, opinion mining intends to identify attributes of the object on which opinion have been given, in each of the document t ϵ T and to find orientation of the comments i.e. whether the comments are positive or negative.

Figure 4 shows different terms that used interchangeably for opinion mining [9]

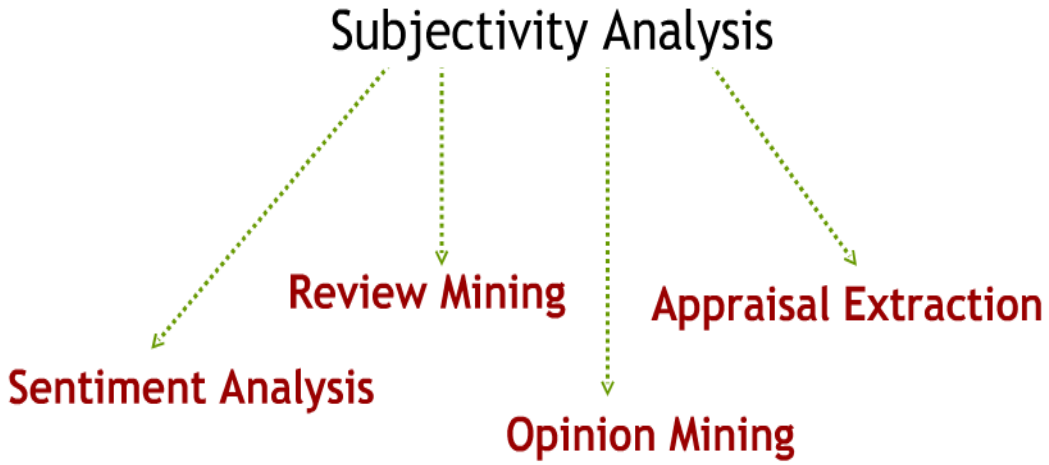

Figure 4: Synonyms of Opinion mining

3.1 Scientific Fundamentals

3.1.1 Model of Opinion Mining

As people are free to give their opinions on anything, e.g., they buy a product and then they express their views on products' features in various forums. The term 'object' is used for the entity on which comments have been given.

- **Definition (object):** An object A is an entity. It is related to a pair, A: (C, R), where C is the components and sub-components of A, and R is the attributes of A. Each component can have its own sub-components and attributes.

"Features" can refer to either components or attributes. It is also commonly used for objects. Let us consider a document t, which contains opinions on an object A. Generally, t is composed of sentences t = (s_1, s_2, s_3… s_n).

- **Definition (opinion passage on a feature):** Opinion on a particular feature f of an object A, extracted from a document t, is a group of sentences in t that contain some opinion on f.

A single sentence may express opinions on several features of a product, e.g., "The picture quality of this camera is good, but the battery life is short".

- **Definition (opinion holder):** The person giving his/her opinion on something is the holder of the opinion.
- **Definition (semantic orientation/sentiment classification of an opinion):** The semantic orientation of an opinion on a feature f states whether the opinion is positive, negative or neutral. This classification can be done at sentence level i.e. whether a sentence contains a positive opinion on a feature of an object or it may contain negative opinion on it.

Fig 5 shows the opinion mining model [9]

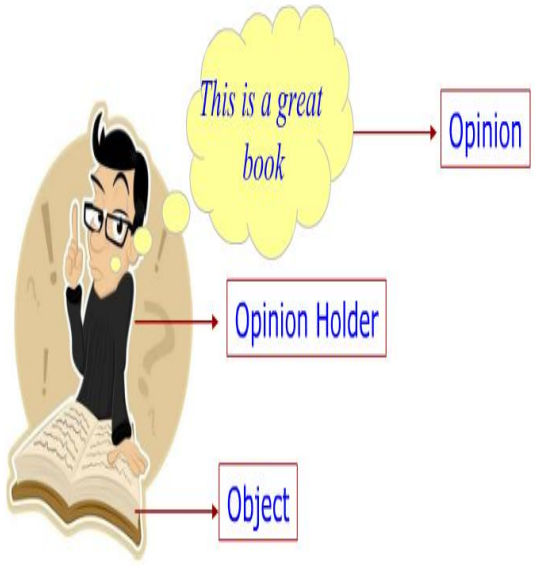

Figure 5: Model of Opinion mining

Putting things together, a model for an object and a set of opinions on the features of the object can be defined, which is called the feature-based opinion mining model.

3.1.2 Model of Feature-Based Opinion Mining

An object A can be represented with a set of features, F = {f_1, f_2… f_n}, which includes the object itself. Each feature fi ϵ F can be expressed with a finite set of words or phrases Wi which are synonyms. That is, there is a set of corresponding synonym sets W = {W_1, W_2… W_n} for n features. An opinion holder comments for each feature f_i to describe the feature by choosing a word from W_k, and then gives opinion on f_i that can be positive, negative or neutral. In a document, Opinion mining is used to extract useful information (sentiments of opinions) from a given document t.

3.1.3 Mining output

Given an evaluative document t having opinions on an object A, the result is a set of quadruples. Every Quadruple is represented by (H, A, f, S), where H being the opinion holder, A being the object, f being feature of the object A and S being semantic orientation of the opinion on feature f in document t.

3.1.4 Opinion Summary

There are several ways to utilize the results of opinion mining. One way is to represent a summary of opinions on features of the objects. This is explained with an example in following section [4, 23]

3.1.5 Mining comparative and superlative (regular) opinion sentences

A person who purchased a certain product can express direct opinion on the features of the object and give comments like "Feature1 of this product is great"."Feature2 of this product is so so". Feature 3 of this product is bogus. Such opinions are direct or regular opinions. In this case, during opinion mining, objects, their features and the orientation of the opinion are to be identified. Such opinions are expressed by using third form of adjectives/adverbs e.g. "Feature4 is best", "Feature4 is worst".

Some people express their views on products in a comparative way e.g. Feature1 of product A is 'better' than Feature 1 of product B. Thus comparisons are made using comparison words and second form of adjectives/adverbs e.g. 'better' in the last sentence. Comparisons are related to but are also different from direct opinions. In such text it is to be identified that what objects are being compared, which of their featured are being compared and which objects are given preference by their opinion holders. [4, 10, 11]

In figure 6, a summary of opinions on features of camera is shown. Here, "CAMERA" represents the camera itself .125 reviews expressed positive opinions on the camera and 7 reviews expressed negative opinions on the camera. "Picture quality" and "size" are two product features. 123 reviews expressed positive opinions on the picture quality, and only 6 reviews expressed negative opinions. The <individual review sentences> points to the specific sentences and/or the whole reviews that give the positive or negative comments about the feature. With such a summary, the user can easily see how existing customers feel about the digital camera. If he/she is very interested in a particular feature, he/she can drill down by following the <individual review sentences> link to see why existing customers like it and/or dislike it. Such sentences are positive or negative or even neutral.

Digital_camera_1:

CAMERA:
Positive: 125 <individual review sentences>
Negative: 7 <individual review sentences>
Feature: **picture quality**
Positive: 123 <individual review sentences>
Negative: 6 <individual review sentences>
Feature: **size**
Positive: 82 <individual review sentences>
Negative: 10 <individual review sentences>
...

Figure 6: Positive and negative opinion on camera featutes

Another way of summary of opinions gathered from multiple reviews on a particular camera is shown in Table 1. These are **regular** opinions marking features of a product as good or bad. The object features and number of positive and negative reviews on these features are listed in the table below.

Table 1 . Number of Positive and negative reviews on camera featutes

Product feature	Reviews (positive)	Reviews (negative)
Camera	125	7
Picture Quality	123	6
Size	82	10

This summary can be easily visualized using a bar chart [4, 8]. Fig. 7 shows such a chart. In the figure, the extent of the bar above the X-axis shows number of positive opinions on a feature (shown at the top of the bar), and the extent of bar below the X-axis shows the number of negative opinions on the same feature.

Obviously, other visualizations are also possible. For example, one may only show the percentage of positive (or negative) opinions on each feature. Comparing opinion summaries of a few competing objects is even more interesting [4, 8]. Fig. 8 shows a visual comparison of consumer opinions on two competing digital cameras. One can clearly see how consumers view different features of each camera. This is comparative opinions i.e. comparing features of two cameras by different companies e.g. Camera A's picture is better than B's.

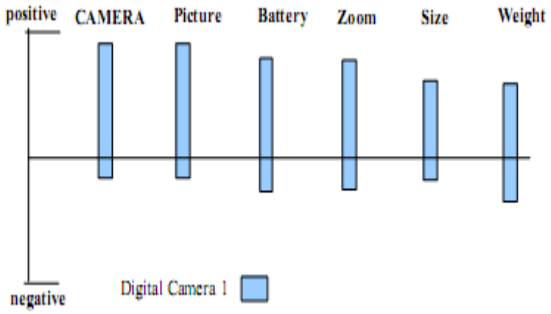

Figure 7: Visual Feature-based summary of opinions on a digital camera

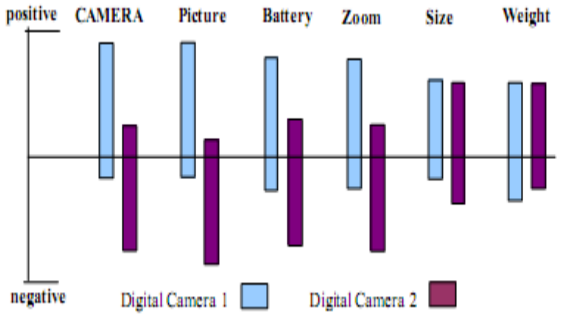

Figure 8: Visual Opinions comparisons of two digital cameras

3.1.6 Sentiment Classification

Sentiment classification has been widely studied in the natural language processing community. It is defined as follows: If a set of evaluative documents T are given, it determines whether each document t ϵ T expresses a positive or negative opinion (or sentiment) on an object. Sentiment classification basically determines the semantic orientation of the opinion expressed on object O in each evaluative document that satisfies the above assumption. Apart from the document-level sentiment classification, classification at the sentence-level is also studied, i.e., classifying each sentence as a subjective or objective sentence and as expressing a positive or negative opinion.

4. CHALLENGES OF OPINION MINING

Since opinion mining is a relatively new filed, thus there are several challenges to be faced. According to Reference [4] current techniques are just primitive for opinions and comparisons identification and extraction. Mainly these challenges are related to the authenticity of the extracted data and the methods used in it. A summary of challenges of opinion mining is as follows:

Reference [27] discusses some issues of opinion mining. A complex example blog to discuss them is considered:

"(1) Yesterday, I bought a Nokia phone and my girlfriend bought a moto phone. (2) We called each other when we got home. (3) The voice on my phone was not clear. (4) The camera was good. (5) My girlfriend said the sound of her phone was clear. (6) I wanted a phone with good voice quality. (7)So I was satisfied and returned the phone to BestBuy yesterday."

Figure 9: A complex example blog

- **Object identification:** In opinion mining, firstly you have to identify the objects in a review on which opinion have been given. This problem is important because without knowing the object on which an opinion has been expressed, the opinion is of little use. However, there is a difference. In for opinion mining, only those objects in the review are to be considered which are in competition to each other. The system thus needs to separate relevant objects and irrelevant objects.
- In the review shown in Figure 9, objects identified are "Nokia phone", "Moto phone", "Bestway". But in the review only "Nokia phone" and "Moto phone" are being compared. So, these are the only relevant entities to be considered for comparison.
- **Feature extraction**: In the review shown in Figure 9, considering the sentence "The voice on my phone was not clear" the object feature is "voice". Reference [26] discusses such a noun based approach in which supervised pattern mining method is suggested. In this technique, frequently used nouns noun phrases as features are identified as features, which are usually genuine features. Many other techniques are also used for extracting information, e.g., conditional random fields (CRF), hidden Markov models (HMM), and many others.
- In the review shown in Figure 9, the features of phone are identified as "camera", "sound" and "voice". It is always a challenge to identify features of the objects. Recently noun based approaches are being used. Verbs can also be the features of an object. But they are difficult to identify.
- **Grouping synonyms:** Different words or phrases can be used to refer to the same feature of the object. So, such words (synonyms) should be identified and grouped together. It is a difficult task to identify these words. A lot of research is required to be done on this issue as it has not been much addressed in the past. To produce a summary similar to the one in Figure 9, it is needed to group synonym features, as people often use different words or phrases to describe the same feature. In this example, "voice" and "sound" both refer to the same feature.
- **Opinion orientation classification:** This task identifies the orientation of opinions i.e. determines whether the opinions on the features are positive, negative or neutral. In the review shown in Figure 9, the opinion on "voice" is negative. Many approaches can be used for this purpose. Usually, lexicon based approach is used as it performs quite well. The lexicon-based approach basically uses opinion words and phrases in a sentence to determine the orientation of an opinion on a feature. A relaxation labeling based approach is also proposed.
- Similarly classifying an opinion as positive, negative or neutral can be a difficult task in opinion mining. A word could be considered positive in one situation and negative in another situation. This can be difficult to calculate as a sentence can be considered negative because of the use of negative words in it.
- The task of extracting the opinion expressed in text is challenging due to different reasons. One of them is that the same word (in particular, adjectives) can have different polarities depending on the context. [14, 16] addresses this problem.
- Existing approaches are based on supervised and unsupervised methods. One of the key issues is to identify opinion words and phrases (e.g., *good*, *bad*, *poor*, *great*), which are instrumental to sentiment analysis. The problem is that there are seemly an

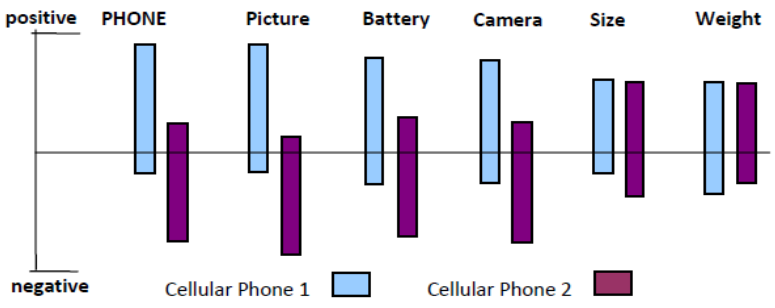

Figure 10: Visual comparison of feature-based opinion summaries of two cellular phones

unlimited number of expressions that people use to express opinions and in different domains they can be significantly different. Even in the same domain, same word depicts different opinions depending upon the context of the text. For example, if a person says "camera has a 'small' size". Then 'small' has a positive orientation because for a camera to have small size is a positive feature as makes it easily portable. But if one says "Camera has 'small' battery life". Then 'small' has a negative orientation because for a camera to have small battery life is negative features as it needs to be replaced frequently. There are still many problems that need to be solved.

- Positive words become negative when used with negations e.g. in the review in example 9 "voice" is the feature and opinion is "clear" which is positive. But as it is stated "voice is not clear". So, negation makes the opinion negative. Hence a context along with word is crucial.

- **Integration:** Integrating the above tasks is also complex because we need to match the five pieces of information in the quintuple (o_j, f_{jk}, oo_{ijkl}, h_i, t_l). That is, the opinion oo_{ijkl} must be given by opinion holder h_i on feature f_{jk} of object o_j at time t_l. To make matters worse, a sentence may not explicitly mention some pieces of information, but they are implied due to pronouns, language conventions, and the context. To deal with these problems, we need to apply NLP techniques in the opinion mining context, e.g., parsing, word sense disambiguation, and coreference resolution. We use coreference resolution as an example to give a glimpse of the issues. For our example blog, to figure out what is "my phone" and what is "her phone" in sentences (3) and (5) is not a simple task. Sentence (4) does not mention any phone and does not have a pronoun. The question is which phone "the camera" belongs to. Coreference resolution is a classic problem in NLP. There is still no accurate solution from the research community.

Reference [4] also discusses some issues:

- **Opinion mining at document level:** Classifying evaluative texts at the document level does not give information about likes and dislikes of the opinion holder.
  An evaluative document usually contains both positive and negative opinions. In the document, Opinion holder does not have all the positive opinions on features of the objects nor does he have disliking for all the features, although the general sentiment on the object may be positive or negative. So, to deal with such a problem, opinion mining at feature level can help to identify both positive and negative opinions in a document.
  So, feature level approach solves this issue. Such kind of problem is addressed in [15] as it discusses agreement or disagreement in forum texts.
- **Selection of opinion oriented sentences. (for comparative and regular)** Textual information can be broadly classified into two main categories[4]:
  - **Facts**
    In past, most of the research has been done on extracting factual information, e.g., information retrieval, Web search, and many other text mining and natural language processing tasks.
  - **Opinions**
    In past no work was done on opinion mining but it is being focused in last few years.

  To separate opinion sentences from others is a basic challenge.

Some other papers also present the issues as:

- **Identify comparison words:** Identify comparison words and whether they are giving positive or negative feedback totally depends on their context. So, it's not an easy job as sometimes good are bad and bad are good.
- **Different People Different Writing Style:** The fact that comments or views entered by people who are different from each other in the way they write, their use of language, abbreviations and their knowledge is a challenge on its own. People also do not express opinion in the same way. One might use certain negative terms in a sentence Text that appears in an online newspaper and that which appears in an online forum is widely different.
  The mining of online forums and discussions is a challenge on its own. Some possible reasons include the use of abbreviations, the entry of comments by different people, who differ in the way they write or in the knowledge of the language they use [15]
  Reference [25] addresses this problem. In this reference some recent works have started to classify product features but they heavily rely on linguistic and natural language processing techniques. However, writing in consumer reviews is usually not formal and does not follow grammatical rules which make language processing approach inappropriate to use. Therefore, the linguistic and language processing approach is not satisfactory. So, a sentiment analysis system is suggested for classifying products features in consumer reviews by means of mining class association rules. The experimental result shows that the content mining approach is much promising than the natural language processing approach.
- **Opinions Change with Time**
  Another challenge lies in the issue of being able to monitor opinions changing with the passage of time. This helps us to observe if a certain product gets improved with time, or people change their opinion about a product and get convinced for it with time. .
  A research work is done to identify how the peoples' mood changes over time in Reference [17].The work done observes blogs where the mood is explicitly specified either by selecting from a predefined list of moods or by entering it as free text.
- **Strength of Opinions**
  Identification of the strength of an opinion is another challenge faced in opinion mining. The strength of an opinion can change as the discussion continues in a forum i.e. arguments used during discussion are strong enough to change the strength

of opinions. To identify strength of opinions SentiWordNet application has been used [20]. Reference [18, 19] addresses this issue.

- **Misleading Opinions due to sarcastic and ironic statements**
  Sarcastic and ironic sentences exist in text. In such a scenario, positive words can have negative sense of usage in a metaphorical manner. So, text in a statement can be hard to identify as sarcastic or ironic which can lead to erroneous orientation and misleading opinion mining. Reference [21] discusses this issue.
- **Sentences with Mixed Views**
  A bigger challenge for opinion mining comes when people express positive and negative review in the same sentence. This is mostly the issue when people are communicating through informal mediums like blogs and forums. People are more likely to combine different opinions in the same sentences. Such sentences can be difficult to parse for opinion mining.
  Sentiment mining or opinion mining is contrasted generally with the traditional fact-based text mining. Text mining seeks to classify documents by topics while opinion mining generalizes text across many domains and users. Strength of a feeling, degree of positivity and similar factors can be of potential importance in opinion mining. This issue is presented in Reference [22] and it suggests a technique to separate positive and negative
- **Misleading Opinions due to spam opinion**
  Reference [24] refers to the issue of dishonest opinions/reviews that intend to affect opinion mining about a product or service. Detecting such opinions is important for practical utilization of opinion mining. Semantic coverage may be useful feature for detecting spam. Spam exists as:
    - Repeating of important terms
    - Dumping of many unrelated terms

Spam opinion and biased opinions are given by people intentionally to affect opinion mining.

5. CONCLUSIONS

Opinions are very important for anyone who is going to make a decision. Web mining has emerged in recent years as an attractive technology to individuals and corporations to know others' opinions.

Opinion mining is helpful for individuals when they want to buy a product and they can decide which product to buy, by studying the summarized opinions instead of studying long reviews and making summary their selves.

Opinion mining is equally important for companies and helps them to know what customers think about their products. Therefore companies can take decisions about their products based on customers' opinion. Thus companies can modify their products according to customers' opinions in a better and faster way. Thus, companies can establish better customer relationship by giving them exactly what they need. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements.

In this paper we have discussed the issues which are faced to do opinion mining from web data and the related work that has been done to deal with these issues.

6. FUTURE WORK

The summarized challenges in this paper have been dealt by several approaches as discussed till now but still there is a great space of improvement in this area. There is need of such an approach which addresses all the challenges of opinion mining simultaneously.

7. DEDICATION

This Research Paper is lovingly dedicated to author's respective parents who have been constant source of inspiration for her. They have given her the drive and discipline to tackle any task with enthusiasm and determination.

8. ACKNOWLEDGEMENT

The author would like to thank Lt. Col. Farooque Azam, whose encouragement, guidance and support from the initial to the final level enabled to develop an understanding of the subject.

9. REFERENCES

[1] Z. M. Ma.2005.Databases Modeling of Engineering Information, Northeastern University, China

[2] M. Mehdi Owrang O.2007.Discovering Quality Knowledge from Relational Databases, American University, USA.

[3] Bing Liu.2011.Sentiment Analysis Tutorial - Given at AAAI-2011, San Francisco, USA.

[4] Bing Liu.2007.Opinion Mining, Department of Computer Science University of Illinois at Chicago.

[5] Andreas Auinger, Martin Fischer.2008.Mining consumers' opinions on the web.

[6] Qingyu Zhang and Richard S. Segall .2008.Web Mining: a Survey of Current Research, Techniques, and Software.

[7] Lita van Wel and Lamber Royakkers.2004. Ethical issues in web data mining, Department of Philosophy and Ethics of Technolog.

[8] Liu, B., Hu, M. and Cheng, J. 2005. Opinion Observer. Analyzing and Comparing Opinions on the Web. Proceedings of International World Wide Web Conference (WWW'05).

[9] Ganesan, K. A., and H. D. Kim. 2008. Opinion Mining - A Short Tutorial (Talk) , University of Illinois at Urbana Champaign.

[10] Ganapathibhotla, G. and Liu, B. 2008.Identifying Preferred Entities in Comparative Sentences. To appear in Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08).

[11] Jindal, N. and Liu, B. Mining.2006.Comparative Sentences and Relations. Proceedings of National Conference on Artificial Intelligence (AAAI'06).

[12] Ana Azevedo AND M F Santos.2008.Kdd, Semma and Crisp-Dm. A Parallel Overview

[13] H D U K Kim, K. Ganesan, P Sondhi, C Zhai. 2011. Comprehensive Review of Opinion Summarization.

[14] Alexandra Balahur and Andrés Montoy.2010.OpAL. Applying Opinion Mining Techniques for the Disambiguation of Sentiment Ambiguous Adjectives in SemEval-2 Task 18.

[15] Anna Stavrianou and Jean-Hugues Chauchat.2008. Opinion Mining Issues and Agreement Identification in Forum Texts.

[16] Andrea Esuli. 2008. Automatic Generation of Lexical Resources for Opinion Mining. Models, Algorithms and Applications.

[17] Krisztian Balog, Maarten de Rijke. 2006.Decomposing Bloggers' Moods - Towards a Time Series Analysis of Moods in the Blogosphere

[18] Marina Sokolova, Guy Lapalme.2008.Verbs Speak Loud. Verb Categories in Learning Polarity and Strength of Opinions.

[19] Animesh Kar, Deba Prasad Mandal.2011.Finding Opinion Strength Using Fuzzy Logic on Web Reviews.

[20] B. Ohana, B. Tierney.2011.Opinion Mining with SentiWordNet.

[21] Zhongwu Zhai, Bing Liu.2011.Identifying Evaluative Sentences in Online Discussions.

[22] Hyun Duk Kim, ChengXiang Zhai.2009.Generating comparative summaries of contradictory opinions in text.

[23] Murthy Ganapathibhotla, Bing Liu.2008.Mining opinions in comparative sentences.

[24] Bo Pang and Lillian Lee.2008. Opinion Mining and Sentiment Analysis, Department of Computer Science University of Illinois at Chicago.

[25] Christopher C. Yang, Y. C. Wong.2008.Mining consumer opinions from the Web.

[26] South Morgan Street, Bing Liu. 2011. Identifying Noun Product Features that Imply Opinions.

[27] Bing Liu. 2010. Sentiment Analysis: A Multi-Faceted Problem.