



Customer Reviews and Analysis using Opinion Mining Adaptive Algorithm

Palla Pavankumar¹·Tippani Gayathri²· Kathi Venkataramana³

Assistant Professor, Dept. of Computer Engineering, Keshav Memorial Institute of Technology, Narayanaguda,
Hyderabad, India¹

Assistant Professor, Dept. of Computer Engineering, Keshav Memorial Institute of Technology, Narayanaguda,
Hyderabad, India²

Assistant Professor, Dept. of Computer Science & Engineering, Geethanjali Institute of Science & Technology,
Nellore, India³

ABSTRACT: The important part to gather the information is always seems as what the people think. The growing availability of opinion rich resources like online review sites and blogs arises as people can easily seek out and understand the opinions of others. Users express their views and opinions regarding products and services. These opinions are subjective information which represents user's sentiments, feelings or appraisal related to the same. The concept of opinion is very broad. In this paper we propose an opinion mining adaptive algorithm (OMAA) focuses on the Classification reviews on customer reviews that conveys user's opinion i.e. positive or negative at various levels. The OMAA is to predicting opinions enable us, to extract the sentiments from the web for online customer's preferences, which could prove valuable for marketing research. Much of the research work had been done on the processing of opinions or sentiments recently because opinions are so important that whenever we need to make a decision we want to know others' opinions. This opinion is not only important for a user but is also useful for predicting the online marketing purchases and improving impacts on products and sales.

KEYWORDS: Opinion Mining, Sentiment Analysis, Mining Classifications, Web Mining

I. INTRODUCTION

We Opinion mining is a type of natural language processing for tracking the attitudes, feelings or appraisal of the public about particular topic, product or services. Textual information in the entire world is of two types: facts and opinions. The facts are the objective expressions which describe the entities, events and properties whereas the opinion is the subjective expression which describes people's opinions, emotions and sentiments towards entities and their properties. The current search engine searches for facts because they assume the facts are true and can be expressed with keywords. But these search engines do not find the opinions because opinions or sentiments are very difficult to express by keywords and that is why there ranking strategy are not appropriate for opinion retrieval. Now, the web has significantly changed its way that people comments their views and opinion on any product and services. User can post their comments on any internet forums, review sites, blogs and discussion group which are commonly known as user generated content which contains the important information. This online word- of- mouth behaviour represents new and considerable sources of information and their applications. These online comments are not limited to our friend circle but it is also expanded on a global or web scale. Today, if the user wants the views on a particular product he/she has no longer limited to ask their friends because they got the opinion for that product on the internet through various reviews or comments. In the same if the organization wants the opinion of their products and services they use these user generated contents from the web for the review or comments of their products and services.

The art Opinion Mining is to recognize the subjectivity and objectivity of a text and further classify the opinion orientation of subjective text. In short we say that Opinion Mining[19] is an automated extraction of subjective content from text and identifying the orientation such as positive or negative in that text. It aims to explore feelings of a person

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

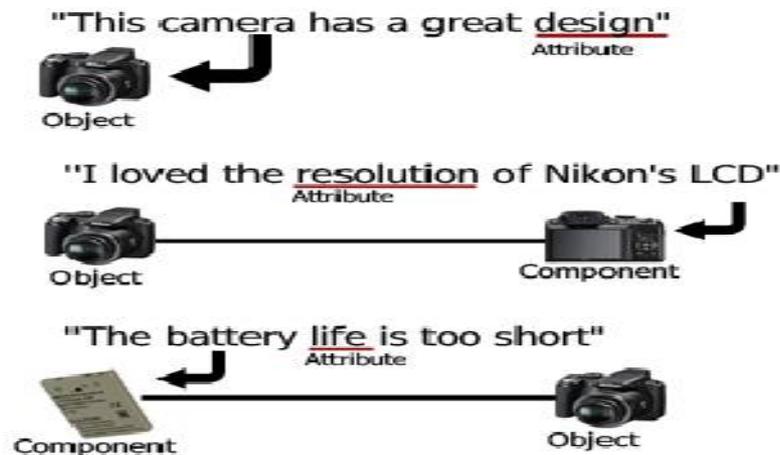
who write the text. It used Natural Language Processing and Machine Learning ethics to determine opinion in the text. The evaluation of opinion can be done in two ways:

- Direct opinion, gives positive or negative opinion about the object directly. For example, “*The picture quality of this camera is poor*” expresses a direct opinion.
- Comparison means to compare the object with some other similar objects. For example, “*The picture quality of camera-y is better than that of Camera-x.*” expresses a comparison.

1.1 Defining Opinion Components in a Opinion Mining Context

The definitions used in this section were proposed in [9], and they summarize important elements that compose an opinion. Some of these definitions are just a natural observation of elements present in opinions, while others refer to definition of problems addressed in [9]. For this reason some of these definitions may not apply to other works as they can differ a lot from their goals as well as from the strategies they employ to achieve them.

1.1 Defining Opinion Components in a Opinion Mining Context



II. RELATED WORK

One of the reasons why some search engines are so successful on the Internet are their commitment to quality services, especially with regard to the processing speed of users queries. In today's Internet, with billions of available pages, the lack of mechanisms that provide a response in a short amount of time would leave the system incompatible with the standards of the modern times (more data must be processed with even more stringent constraints with respect to time). The main names behind search engines (e.g. Google, Yahoo, Bing, etc.) achieve this high standard of service mainly due to their indexing technique combined with high end infrastructure comprised of several hundred of clusters highly optimized for jobs demanding large capacity of computational work. A ranking algorithm provides a hierarchical level of more important documents, thereby providing an initial clue to the user about where the desired information is more likely to be. During a search operation the following interactions are performed, as shown in figure 2.1: (1) A query is submitted by the user. (2) The user query is checked to ensure that it is ready to be used by the retrieval system. This could be achieved through simple tasks such as removing stopwords, reducing the words to roots (stemming) and checking the spelling. (3) The query is checked against the available indexes in order to retrieve the documents that contain some of the query terms. Afterwards, a ranking algorithm is applied to the set of found documents which are finally presented to the user (the most relevant documents appear at the beginning of this list). (4) The user receives the response and accesses the matching documents from a result list.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

2.1 Web Content Mining

Web content mining is very important as it deals directly with information. The goal is to mine content from web documents in order to build knowledge from it. This knowledge can be either hidden or somehow simply difficult to be analyzed in a straightforward way. Next subsection will introduce an important field of Web Content Mining called Opinion Mining.

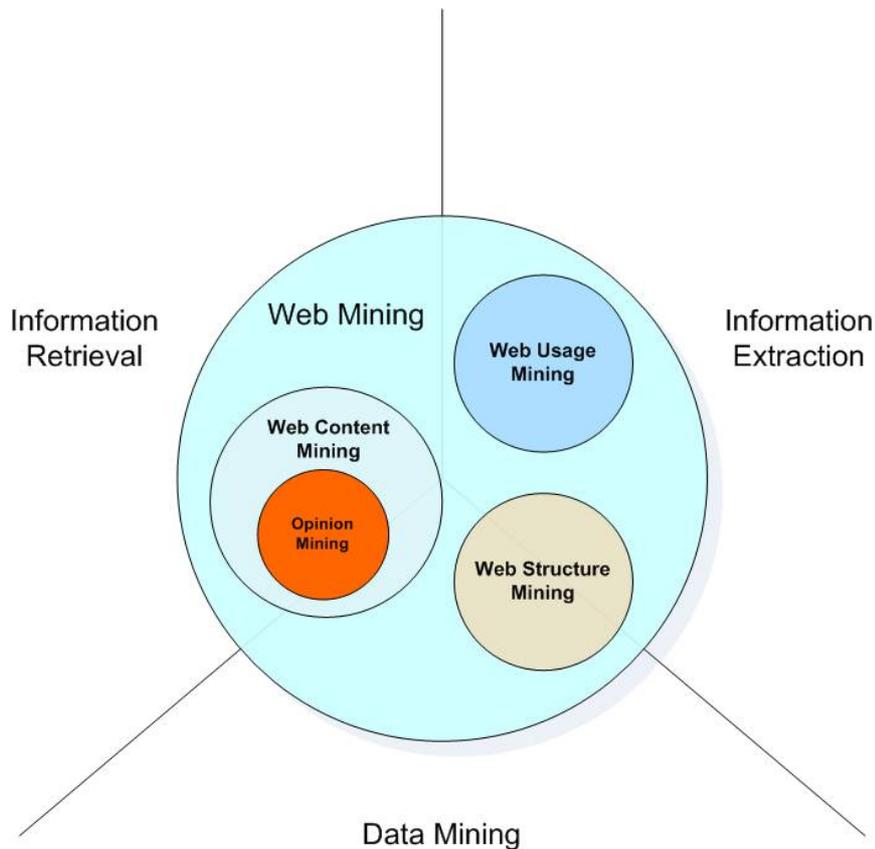


Fig 2.1: Information retrieval systems.

2.3 Opinion Mining

Opinion Mining is a field of Web Content Mining that aims to find valuable information out of users opinions. Mining opinions on the web is a fairly new subject, and its importance has grown significantly mainly due to the fast growth of e-commerce, blogs and forums. With the high profits of e-commerce increasing year after year many people had changed the habit of going to a shop for the comfortable virtual shopping. Eitherways, searching for useful information on users opinions before purchasing a product, became a common practice for many people. A major problem however, is finding the desired information on them. It is not difficult to find web sites with thousands of reviews for a single product, and thus finding an useful information among them can be a very difficult task. For example, a new customer may be interested in reviews that talk about many features of a certain camera. However, old customers of a specific brand, may be interested only in what people have to say about the auto focus function of a newly released model. From the business perspective getting important information out of opinions can represent a good source of advertisement or product feedback. For example, a website specialized on electronics reviews could place advertisements on their pages based on consumer opinions. For instance, if the majority of users express negative

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

opinions about a given product, the web site could place ads from an alternative product from a competitor. Also, manufacturers can get the feedback of their products to improve their products or services.

III. SYSTEM ARCHITECTURE

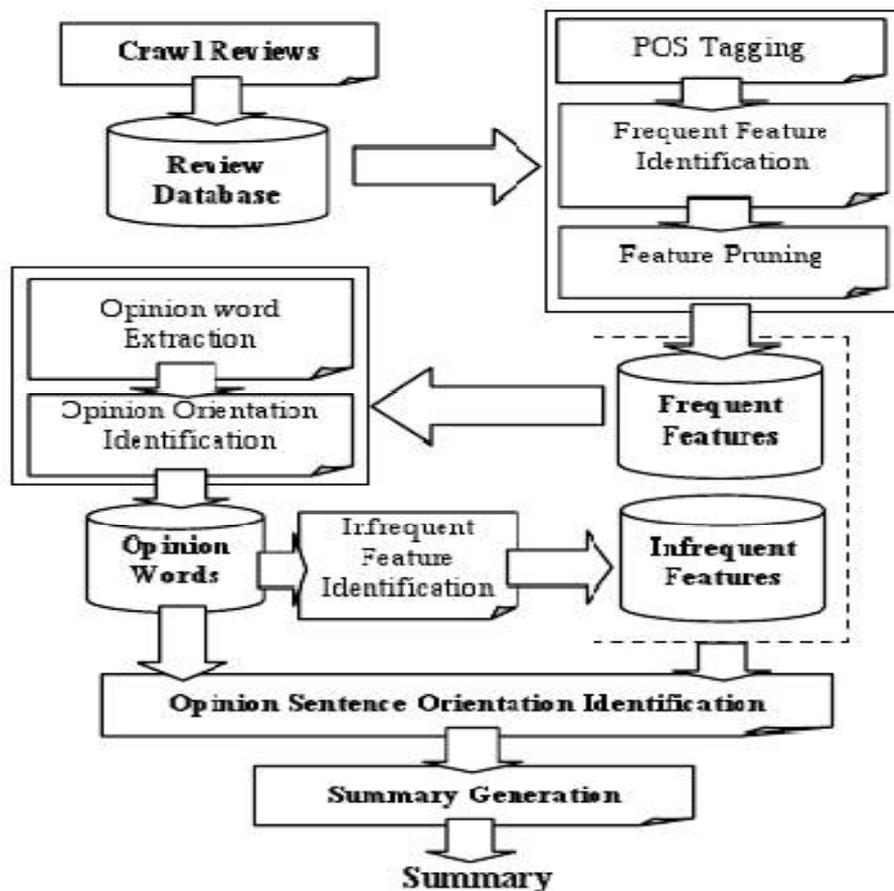


Figure 3: Architecture of an opinion mining system

The system counts with a crawling module, which first downloads all the reviews and stores them in the database. After that a POS tagger tags all the reviews which will work as hooks for the mining part responsible for finding frequent features. This step is skipped by some systems which employ manual feature annotation as in [21], where ontologies were used to annotate movies features manually. Next, with the tagged sentences and features identified, opinion words are extracted and their semantic orientation is identified with the help of WordNet. Now with opinion words identified and extracted, the system identifies infrequent features.

IV. ANALYSIS

4.1 Feature Identification

Feature identification is the process used to deduce possible product features out of the tagged texts generated by the last step. Both [15] and [9] use some heuristics to narrow words that are more likely to be a feature inside a sentence. Normally, the part-of-speech responsible for giving names to entities of the real world are nouns, in this case a noun gives name to the product and its features (i.e zoom, battery life, image quality, etc.). In these works, they define two categories of features, frequent features and infrequent features. In [21] Ontology Based Opinion



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

"Mining for Movie Reviews", an ontology based approach was used to extract features from opinions. In their work, they experimented with movies reviews, where they identified sentences containing the ontology terminologies.

Here it is important to differentiate between the two approaches with their pros and cons. In [15] and [9], feature identification is performed automatically. The great advantage of this method is to perform the whole process automatically, with minimal human intervention. The biggest downside however is that the output (the frequent features) depends a lot on the number of opinions being analyzed. Also, there is no guarantee that a frequent feature found by the system is actually a real feature. In [21] and other works where features were manually annotated, the advantage is that the system will always identify real features, being frequent or not. This will only depend on the correctness of the previously made annotation. The great downside however, is that a great number of annotations have to be made. They may not be only specific to categories (such as digital cameras, video games, cellular phones), but they could be even more specific such as models of a specific brand (Nikon P90, Nikon D5000, etc). That would make the annotation of features a very hard work. Also, people may comment on the lack of features of a given product, or they may use different words to refer to the same feature which a system with manual annotated features will fail to recognize. Given the short comparison between the different approaches above, the method explored in [15] and [9] will be further discussed, as it needs a minimal human intervention to perform its task reasonably well which could be later improved by other methods.

4.2 Frequent Features Identification

In, [15] and [9] the proposed system extract only nouns or noun phrases (explicit possible features) from the text. In this step, the extracted nouns are called candidate features.

Then an association mining algorithm finds all the frequent itemsets, which are the set of frequent features (those ones that many users write about). The idea behind this technique is that features that appear on many opinions have more chance to be relevant, and therefore, more likely to be actually a real product feature. The Apriori algorithm [2] was used to generate the set of frequent itemsets. However, for this task there was no need of finding association rules among items, therefore only the part of the algorithm that finds frequent itemsets was interesting for these works. In [9] and [15], a minimum support of 1% was used.

4.2.1 Infrequent Features Identification

A very simple heuristic was used in [15] to discover possible infrequent features (the ones referenced by a small number of people), as the association mining is unable to identify them.

Example:

- I. "The pictures are absolutely amazing."
- II. "The software that comes with it is amazing."

In the above example, the two sentences have one opinion word in common: amazing. Because an opinion word could be used to describe more than one object, these opinion words are used to look for features which couldn't be found in the step described on

```
for each sentence in the review database
  if (it contains no frequent feature but one or more opinion
      words)
    { find the nearest noun/noun phrase around the opinion
      word. The noun/noun phrase is stored in the feature
      set as an infrequent feature. }
```

Figure 4.2.1: Infrequent feature extraction

4.3 Opinion Sentiment Analysis

Sentiment classification or sentiment analysis is an area of study that aims to classify the sentiment encoded by texts as shown in the following example:

Example:

- I. The girl is angry → negative



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

II. The sun is absolutely beautiful today → positive

The word sentiment is a synonym for polarity and both are widely used to describe the orientation of texts, sentences and words as in example 4. This work will deal with sentiment classification of texts represented by users' opinions, therefore the name opinion sentiment. The analysis of sentiment can be performed on several levels of granularity (words, sentences, texts and documents). For many applications, classifying the sentiment of documents as a whole is sufficient, for others a finer level of granularity might be necessary. The work done by [18] and [22] classifies each user opinion as a whole. In [15] and [3] sentiment classification is done at the sentence level. In [9], each feature within an opinion has a sentiment associated. The reason why this last approach is preferable than the others, is easy to realize through a simple observation. To illustrate it, think about a specialized web site for cameras, where customers can write their opinions about a certain product, as illustrated in figure 3.4. For example the sentence: "I like my camera and the 24x zoom, but I think the battery life is too short". Here it is easy to realize that the sentence is "more positive than negative", however that would still hide one negative aspect of the camera under discussion. This may represent a very important piece of information, which can be obscured by classifying the whole sentence as positive. For this reason the method explored by [9] achieves an optimal granularity level as it treats each attribute of an OuD with the necessary details.

4.3.1 Determining Sentiment of Opinions at the Sentence Level

An opinion can be analyzed at different levels of granularity. In this work, the approaches used to analyze the opinion as a whole are going to be discarded as in [18] and [22]. Therefore, the two remaining approaches with respect to their granularity, are going to be exposed with more details. Figure 4.2.1 shows a pseudocode which aims to find the sentiment of opinions at the sentence level. The next section, will analyze the sentiment of an opinion at the feature level, as proposed by [9], which is of greater importance for this work.

4.3.2 Negation Rules

A negation word such as no, not and never and also some words that follow patterns such as "stop" + "vb-ing", "quit" + "vb-ing" and "cease" + "vb-ing" change the orientation of opinion words in the following way:

- I. Negation Negative → Positive
- II. Negation Positive → Negative
- III. Negation Neutral → Negative

Some examples for each negation rule defined above:

- I. "no problem"
- II. "not good"
- III. "does not work"

4.3.4 Determining Opinion Sentiment at the Feature Level

In [9], after identifying all the opinion words for a given feature, the system calculates the opinion orientation for each feature using the following equation:

$$\text{Score}(f) = \sum_{wi:wi \in S \wedge wi \in V}^n \frac{wi.SO}{dis(wi, f)}$$

- S is a sentence with a set of features
- wi is an opinion word
- V is the set of all opinion words (including idioms)
- $wi.SO$ is the semantic orientation of opinion word wi
- $dis(wi, f)$ is the distance between feature f and opinion word wi in the sentence S

For a sentence s that contains a set of features and for each feature f the above orientation score is computed. A positive word is assigned the orientation +1, and a negative one is assigned -1. In the above equation wi is an opinion

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

word, V is the set of all opinion words (including idioms) and s is the sentence that contains the feature f , $dis(w_i; f)$ is the distance between feature f and opinion word w_i in the sentence s . The reason for the multiplicative inverse in the formula is to give low weights to opinion words that are far away from the feature f . The pseudo code of figure 3 was used to find the opinion orientation at the feature level.

OPINION MINING ADAPTIVE ALGORITHM (OMAA)

```

for each sentence  $S_i$  that contains a set of features do
features = features contained in  $S_i$ 
for each feature  $f_i$  in features do
     $Adaption = 0$ ;
if feature  $f_i$  is in the "but" clause then
     $Adaption =$  apply the "but" clause rule
else remove "but" clause from  $S_i$  if it exists;
for each unmarked opinion word  $OW$  in  $S_i$  do
     $Adaption += word\ Adaption(ow, f_i, S_i)$ ;
endfor
endif
if  $Adaption > 0$  then
 $f_i$ 's  $Adaption$  in  $S_i = 1$ 
else if  $Adaption$  in  $S_i < 0$  then
 $f_i$ 's  $Adaption$  in  $S_i = -1$ 
else
 $f_i$ 's  $Adaption$  in  $S_i = 0$ 
endif
endif

```

V. IMPLEMENTATION

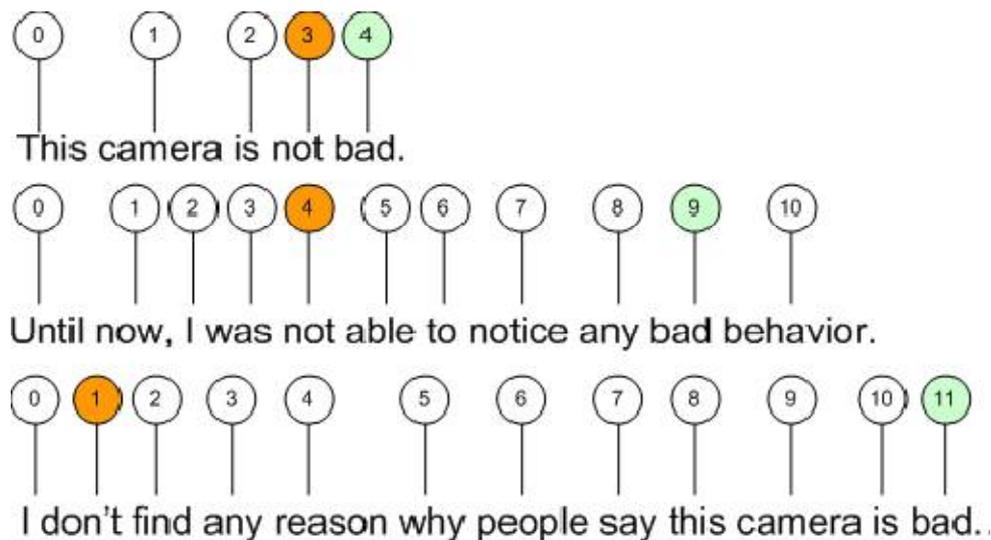


Figure 5.1: Sentences with negation words

To evaluate the effectiveness of the sentiment classification algorithm, the orientation associated with each feature in a sentence was analyzed manually in order to achieve a high degree of confidence. The result of this analysis is presented on the graph of figure 6.1 and table 6.2. A correctly classified opinion is either a

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

negative or positive opinion for a given feature, which was correctly identified by the system. All the features were identified automatically, and only the sentences with real features (frequent features) were analyzed. Sentences with candidate features were discarded. The effectiveness of the automatic feature identification algorithm will be discussed later on section 7.3. It is important to consider the complexity of judging (even for a human) whether an opinion is positive or negative. There are many cases where this analysis is pretty hard. ObjectSpace is a Ruby module used for memory management, which is disabled by default in JRuby. One of POECS libraries use this feature which can be enabled by passing the parameters -X+O. It has been reported by the JRuby developing team that ObjectSpace can create a substantial overhead and thus affecting the performance of the system.

Table 5.1: Effectiveness of sentiment classification

Product	Correctly Classified Opinions	Opinion Sentences
Ipod Touch 8GB	79%	673
Nikon D5000	82%	452
Nikon P90	84%	273
Xbox 360	73%	181

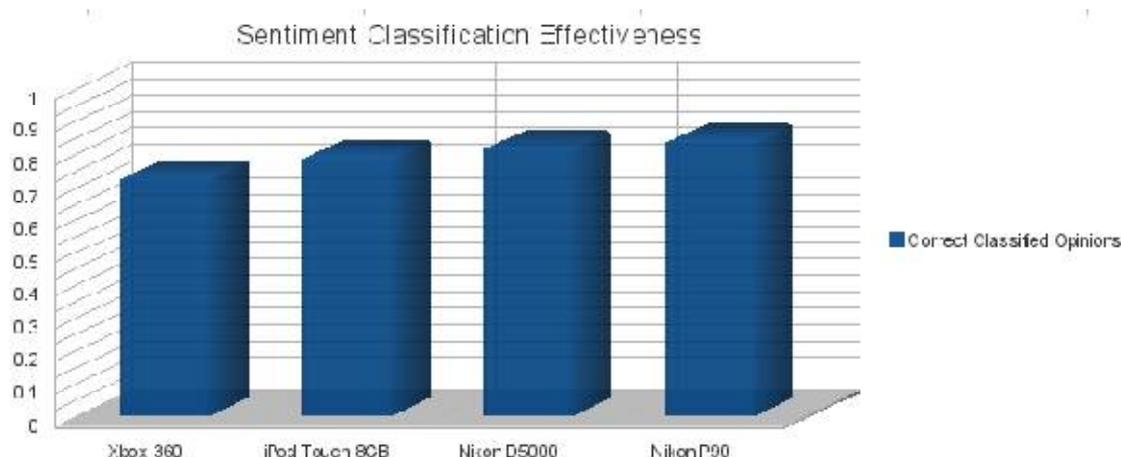


Figure 5.1: Effectiveness of opinion sentiment classification

IV. EVALUATION

Straightforward, especially when explicit opinion words are used. However, in some cases, in order to understand whether an opinion is negative or positive, a domain specific knowledge is needed. Example 1 illustrates some of these cases:

Example 1:

- I. "The iPod battery lasts for 5 hours with music and Wi-Fi turned on."
- II. "The device heats very fast."

In sentence I, it is difficult to know whether a battery lasting for 5 hours under the described conditions is positive or negative. In sentence II, heating fast depends a lot of which device is being analyzed. For instance getting hot fast could represent negative behaviour of an electronic device such as video games or notebooks. However, if the OuD is a portable heater or an electric oven, heating fast should be a positive behaviour.

There are also situations where wrong parts-of-speech are assigned to words and since the POS tagging operation is the foundation of the whole mining process, a wrong tagging classification would directly affect the mining process, as shown in example 2:



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Example 2:

- I. "This battery is really GOOD."
- II. "[. . .] hard drive of 120 GB which comes with the device."

In sentence I, the word GOOD is written in upcase, a common practice among internet users, which aims to emphasize sentiments. The POS tagger however interprets this word as belonging to the group of proper nouns. Therefore, the mining algorithm will fail to recognize it as an opinion word. In this work, such cases were not covered since they are not common and hence addressing them would produce a great delay for the mining tasks. A snippet of a customer opinion for Xbox 360 is shown in sentence II. It was detected that the POS tagger would fail to recognize hard drive as a noun group (feature). Unlike, the POS tagger interprets the word as follows: hard → adjective and drive → noun. Thus, hard was taken by the mining algorithm as an opinion word.

VI. CONCLUSION & FUTURE WORK

Opinions are a unique type of information which are different from facts. The methods for content classification based on ranking (like those used by search engines) are not effective or simply do not accurately depict reality, as one opinion is different from multiple opinions.

During the evaluation of POECS it was possible to see that it is feasible and reliable to build a system capable of classifying and organizing opinions through the so-called feature-based summary, which resumes the most relevant information for users.

However, it is undeniable that a great number of opinions are difficult to classify due to the complexity of the human language. Evaluation also showed that the system can be more effective when domain-specific, using the help of manual annotations to treat common exceptions. A system can therefore combine multiple approaches with the intelligence of automatic algorithms and manual annotations in order to provide a high degree of accuracy.

Many studies including this one are striving to discover patterns of use of language that can be reused, widespread and processed by computers. While many annotations are still required, they do not offer the necessary flexibility and often turn out to be very specific to a domain. There are complex cases as the one presented during evaluation (which the system fails to correctly classify) which remain as challenges for new algorithms which are capable of solving disambiguation problems.

As discussed, when one is commenting about a product, a word can have different meanings or can refer to different objects. Solving this problem is a good challenge for new algorithms in Opinion Mining. Also, understanding the semantics of text in a more intelligent way is necessary. A good direction would be methods which have global knowledge of opinions dependent on complex contexts, which can use this information later to help solving context problems in any local analysis (at the sentence or word level).

REFERENCES

- [1] Nidhi Mishra C.K.Jha, PhD Classification of Opinion Mining Techniques *International Journal of Computer Applications* (0975 – 8887) 2012
- [2] J. Handl, J. Knowles and M. Dorigo Ant-based clustering: a comparative study of its relative performance with respect to k -means, average link and 1d-som 2016
- [3] Horacio Sagion Kalina Bontcheva Christian Leibold Adam Funk, Yaoyong Li. Opinion analysis for business intelligence applications. OBI'08, 2009.
- [4] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. 1994.
- [5] Oren Etzioni Ana-Maria Popescu. Extracting product features and opinions from reviews. EMNLP-05, 2005.34
- [6] Vadim Tkachenko Jeremy Zawodny D. Arjen Lentz Derek J. Balling Baron Schwartz, Peter Zaitsev. High Performance MySQL: Optimization, Backups, Replication, and More. O'Reilly Media, 2008.
- [7] Liu Bing. Web Data Mining. Springer, 2007.
- [8] Max Bramer. Principles of Data Mining. Springer, 2007.21
- [9] Brian Clifton. Advanced Web Metrics with Google Analytics. Sybex, 2 edition, 2010.
- [10] Yukihiro Matsumoto David Flanagan. The Ruby Programming Language. O'Reilly, 2008.
- [11] Philip S. Yu Ding Xiaowen, Liu Bing. A holistic lexicon based approach to opinion mining. WSDM'08, 2008.vii,27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 58, 59, 87
- [12] Brad Ediger. Advanced Rails. O G Pops, 2008.
- [13] David A. Freedman. Statistical Models: Theory and Practice. Cambridge University Press, 2 edition, 2009.
- [14] Tom Funk. Web 2.0 and Beyond: Understanding the New Online Business Models, Trends, and Technologies. Praeger, 2008.
- [15] Abraham Kandel George Meghabghab. Search Engines, Link Analysis and User's Web Behaviour. Springer, 2008.14



ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

- [16] Lynne Schrum Gwen Solomon. Web 2.0: New Tools, NewSchools. International Society for Technology in Education; First Edition edition, 2007.
- [17] Liu Bing Hu Mingqing. Mining and summarizing customer reviews. KDD'04, 2004.vii, 27, 30, 31, 32, 33, 34,35, 38, 59.
- [18] Eibe Frank Ian H. Witten. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann;2 edition, 2005.
- [19] Paul Kimmel. UML Demystified. McGraw-Hill OsborneMedia; 1 edition, 2005. [20] David M. Pennock Kushal Dave, Steve Lawrence.Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.WWW'03, 2003. 34, 36
- [21] Arthur M. Langer. Analysis and Design of InformationSystems.Springer; 3rd edition, 2007.
- [22] David Heinemeier Hansson Leonard Richardson,Sam Ruby. Restful Web Services. O'Reilly Media; edition, 2007.
- [23] Chunping Li Lili Zhao. Ontology based opinion miningfor movie reviews. KSEM 2009, 2009.
- [24] Simon Corston-Oliver Eric Ringger Michael Gamon, Anthony Aue. Pulse: Mining customer opinions from freetext. IDA, 2005.34, 36.

BIOGRAPHY

PallaPavankumaris anAssistant professor in the Computer Science and Engineering Department,Keshav Memorial Institute of Technology, Narayanaguda, Hyderabad. He received Master of Technology (M.Tech) degree in 2012 from SKD college, Gooty, Ananatapur, India. His research interests are Data Mining, Computer Networks (wireless Networks), Image Processing, Algorithms etc.

TippaniGayathriis an Assistant professor in the Computer Science and Engineering Department,Keshav Memorial Institute of Technology, Narayanaguda, Hyderabad. She received Master of Technology (M.Tech) degree in 2014 from JNTU, Hyderabad India. Her research interests are Image Processing, Data Mining, Computer Networks (wireless Networks) etc.

TippaniGayathriis an Assistant professor in the Computer Science and Engineering Department,GeethanjaliInstitute of Science & Technology, Nellore. He received Master of Technology (M.Tech) degree in 2013 from JNTU, Anantapur, India. Her research interests are Image Processing, Data Mining, Computer Networks (wireless Networks) etc.