Semantic Orientation of Sentiment Analysis on Social Media

1Sri Khetwat Saritha, 2Devshriroy
1,2 Department of CSE & IT, Maulana Azad National Institute of Technology
Email: 1skmanit@gmail.com, 2droy.iit@gmail.com

ABSTRACT
User Generated Contents on social media, such as Blogs, Forums, YouTube, Twitter, Facebook and so on contains opinions or sentiments generated by the users about the object, such as product reviews, movie reviews, book reviews etc. The texts in these social media sites are short and generated constantly. These contents are well suited for knowledge discovery. The purpose of this paper is to locate, extract, classify and summarize the customer opinion about the products from the social media site YouTube. The proposed framework determines the semantic orientation of opinion expressed on product features as positive, negative or neutral. The system is also integrated with a visualization module to present feature based summary of user generated contents.

Indexing terms/Keywords
User Generated Contents, Social Media, Opinion Mining, Sentiment analysis, Alchemy API.
INTRODUCTION

A number of review sources are available in the Web where reviews about product can be found, like in E – Commerce sites (Amazon), forums, blogs, and social media site like YouTube, Facebook, Twitter and so on. These sources are very popular among the customers. A landscape of the social media [26] is shown in the figure 1.

Figure 1: Social Media Landscape

Online selling and purchasing of products are increasing day by day. Customers give their views regarding the product under customer’s review generally which includes customer’s natural language. These reviews are known as user generated contents.

Customers generally express their views in natural language. For example “Size of Nokia Lumia is very good and handy”. The feature “size and handy” of the mobile “Nokia Lumia” is highlighted. Mostly a customer expresses opinion about a product as a whole or about its features as shown in figure 2. Such customer reviews helps the company to know and analyze the public or customer opinions about their products to establish future directions for improvement [1].

![Figure 2: Review on Nokia Lumia 800 (example)](image)

Important thing in classifying reviews is the choice of feature set. The feature based opinion extraction is a task related to information extraction, which consists in extracting structured opinions on features of some object from subjective texts [2]. The problem of feature based sentiment extraction is divided into two different tasks [3]:

1. Features of the product about which the reviewer have expressed their opinions are to be identified and extracted.
2. The semantic orientations or the polarity of the opinions are to be determined.

The semantic orientation [4, 5] or polarity of a term indicates the positive or negative implications of that term being used in an opinion. In this paper, a feature based sentiment analysis framework is shown to determine opinions that are expressed on different features of the product.

USER GENERATED CONTENT RESOURCES

User generated contents holds the information about a product. It can be collected from various blogs, review sites, micro blogging and social media sites like YouTube, Twitter, Facebook and so on. This text is unstructured and unmanaged which needs proper arrangement to extract knowledge from it.

1. **Review Sites**

   Review site[6] is a website on which reviews can be posted about businesses, people, product or services. Early review sites included Epinions.com and Amazon.com.
(2) **Blogs**

Due to continuous growth, the number of blogs doubles every 5 months and there are new blogs published every second [7]. A Blog [8] is a discussion or information site published on the World Wide Web and consisting of discrete entries (“posts”) typically displayed in reverse chronological order (the most recent appears first).

(3) **Social Media**

There are many social media sources available on World Wide Web, among them

**YouTube** [9] allows billions of users to discover, upload and share originally created videos. It provides forum for people to connect, inspire others across the globe.

**Twitter** [10] is an online social networking service and micro blogging service. This service connects to latest stories, ideas, opinions and news. It enables its users to send and read text-based messages up to 140 characters, known as “tweets”.

**Flickr** [11] is a photo sharing websites where anyone can upload and tag photos, browse others’ photos, and add comments and annotations. Users can create photo set and collections to manage content and participate in topical groups to cultivate a sense of community.

**OPINION MINING AND SENTIMENT ANALYSIS**

The increasing user generated contents on the internet forced researchers to study the problem of opinion mining. It is a growing research area in natural language processing, computational linguistic, text mining and information retrieval. Opinion mining is a research domain deals with the methods of detecting and extracting opinions and sentiments from the user generated contents.

The definition of opinion mining was proposed by Pang & lee [6]. Opinion is a private state of a person thinking about something [12]. Finding opinion about the product from online sites is a very big task because finding the reaction of the customers from the large review set becomes difficult for the customer if goes manually by reading all reviews, which is time consuming. As these contents are unstructured and are unmanaged. Figure 3 shows the different opinions expressed by the customer about the product and there features.

![Figure 3: Semantic orientation of reviews](image)

**Figure 3: Semantic orientation of reviews**

Opinion mining from user generated contents can be done at document level, sentence level and entity and feature or aspect level. There are two types of opinion [13] – Regular opinions which target at some entities/features and Comparative opinion which make comparison of more than one entity, as shown in figure 4.

![Figure 4: Holding Opinion by comparing with other](image)

**Figure 4: Holding Opinion by comparing with other**
In Figure 5 the negative opinion expressed by the customer on a product is shown.

![Figure 5: Customer stating Negative Opinion on few features](image)

In some reviews the whole text does not represent opinion but only a portion of a review or some sentences include opinion oriented words as shown in figure 6.

![Figure 6: Lengthy review with few Semantic orientations](image)

There are many application areas where sentiment analysis can be used, such as analysis on financial markets, analysis on products, analysis on location (tourism), analysis on elections, analysis on movies and software programs, and so on.

**PREVIOUS WORKS**

Various mining techniques in extracting opinions generated by the user on different entities/features are studied by the researchers and various approaches are given by the researchers.

A heuristic feature based extraction algorithm given in [15] depends on feature terms with their number of occurrence. They use association rule mining based on Apriori algorithm to extract frequent item sets as explicit product features. In this paper we have used Apriori algorithm to count the occurrence of the feature terms in the user generated text. An ontology based opinion mining extraction system given in [16], this system worked semantically. They constructed the ontology manually and updated as new features were added.

An unsupervised information extraction system called OPINES was developed by [17] which mine product reviews to build a model of product features, their evaluation by reviewers and their relative quality across products. A heuristics based feature extraction algorithm is given in [18], they used association rule mining based on the Apriori algorithm to extract frequent item sets as explicit product feature. A rule based approach for feature extraction is given in [19], where they extracts large number of features from the reviews and removes multiple conflicting words in a sentence.

A semantic orientation of phrases was developed by [20] they proposed a PMI – IR algorithm to estimate the semantic orientation of a phrase. Based on POS tagging [21], the author carried out the extraction of frequent features, opinion words and infrequent features through the co-occurrence relationship among them.

In this paper, a framework is designed to determine the semantic orientation of opinion expressed on product features as positive, negative or neutral. We locate, extract, classify and summarize the customer opinion about the products from the social media site YouTube.

**OPINION EXTRACTION SYSTEM**

The framework involves various steps in extracting the opinions from the user generated contents. They are Crawling, Preprocessing and Semantic Orientation.
(4) Crawling

Web Crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion [22]. It is relatively simple automated program, or script, that methodically scans or crawls through Internet pages to create an index of the data [23]. Extraction of UGC done by crawling YouTube based on the request of the user is shown in the figure 7.

Figure 7. Semantic orientation and features of mobile in a review

Figure 8: Extraction of UGC from YouTube

The Videos are searched by sending a query to the YouTube server using product name as the filter. All Video ID are downloaded which have their unique Id number.

For Example:

```
For each unique video id, a request is sent which returns the XML file with comments.  
```

```
http://gdata.youtube.com/feeds/api/videos
```

```
Produce Name | Feature Set | Video ID
-------------|-------------|-----------
Nokia Lumia  | Battery Life| 20XFhQpmh2E 
              |             | msbE0LxpYKs 
              |             | QvnDunLJSPE 
              |             | MJyqURfRJ4w 
              |             | YS9azQLWNAI 
              |             | and so on.
```

```
``
The comments are extracted within the `<comment>` tag. An Algorithm 1 shown below describes the extraction of reviews from YouTube:

### Algorithm 1: Extract Reviews from YouTube.

**Step 1:** Input {product name, feature set}

**Step 2:** Searching for the video from YouTube by sending a query to the YouTube Server URL:

http://gdata.youtube.com/feeds/api/videos

**Step 3:** Extract all Video Ids based on the product name as a Filter.

Output: {VideoID.txt}

**Step 4:** Store all features into a file. {Feature.txt}

**Step 5:** For each VideoID in VideoID.txt perform step 6 to step 8

**Step 6:** Read the VideoID and send each Video ID to URL

http://gdata.youtube.com/feeds/api/video/VIDEOID/comments

**Step 7:** Return XML file with comments

**Step 8:** Read the XML and extract the text within `<comment>` tag

Output: {comment.txt}

#### Preprocessing and Semantic Orientation

While processing the text, stop words are removed and stemming is done. The unwanted symbols are removed using ASCII check for each word. Figure 9 shows the Extraction of semantic concepts of the User Generated Contents.

![Figure 9: Extraction of Filtered UGC based on feature set from YouTube and Semantic Orientation of the comments](image)

After the removing the noise from the reviews, the reviews are separated based on the feature set given by the user. For opinion word identification and semantic orientation AlchemyAPI is used [24]. AlchemyAPI provides easy-to-use mechanisms to identify positive / negative sentiment within any document or web page. The “URLGetTextSentiment” call is utilized to extract positive / negative sentiments. The result obtained in XML form.

```xml
<results>
  <status>REQUEST_STATUS</status>
  <url>REQUESTED_URL</url>
  <language>DOCUMENT_LANGUAGE</language>
  <text>DOCUMENT_TEXT</text>
  <docSentiment>
    <type>SENTIMENT_LABEL</type>
    <score>DOCUMENT_SENTIMENT</score>
    <mixed>SENTIMENT_MIXED</mixed>
  </docSentiment>
</results>
```

![Figure 10: Response format of AlchemyAPI call (XML)](image)
The algorithm shown below shows the extraction system and semantic orientation of the User generated contents.

**Algorithm 2: Preprocessing of Extract Reviews from YouTube.**

Step 1: Process the comment file to remove unwanted symbols using the ASCII check for each word.

Step 2: Identify the keywords of feature file in the comment file.
   Input: {comment.txt, feature.txt}

Step 3: Classify the comments in comment.txt and create a separate file for each feature say battery.txt, camera.txt and so on for the product: mobile phone

Step 4: For each comment file generated in Step 3 perform Step 5 to Step 6

Step 5: Perform the stemming of words in each feature file containing comments, identify the Opinion words
   Input: {battery.txt, camera.txt, . . .}

Step 6: Calculating the Semantic Orientation of the comments, built using Alchemy API.

Step 7: Create a Bar graph with feature on X-axis and the polarity score on Y-axis.

**RESULTS AND DISCUSSION**

Mobiles phones relatively have a small set of features. We annotated the reviews crawled from YouTube based on the request of the user. We extracted around 40 features which are in common from the reputed mobile selling sites. An interface was designed where the user can input the product name and the features set for which he would like to know the opinion of the other customers about the product on given feature set.

![Figure 11: Interface for entering the Product name and required featured set](image)

AlchemyAPI[25] utilizes natural language processing technology and machine learning algorithms to analyze content, extracting semantic meta-data: information about people, places, companies, topics, facts & relationships, authors, languages and more. API endpoints are provided for performing content analysis on Internet-accessible web pages, posted HTML or text content.
We analyzed around 30 mobile phones and their features. We received a good result for the given product and feature set. Figure 6 shows the result for one for the product – Samsung Galaxy.

CONCLUSION

This paper describes the method for extracting the comments from the YouTube based on the user request. The main idea behind this work was to determine the semantic orientation of customer opinion about the products so that one can easily know about the pros and cons of a product and can make decision accordingly.

REFERENCES


[9] www.youtube.com/t/about_youtube


[14] Ion Smeureanu, Cristian Bucur, ”Applying supervised opinion mining techniques on online user reviews”, Informatical Economica vol. 16. No. 2/2012 pp. 81 - 91


[22] en.wikipedia.org/wiki/Web_crawler

[23] www.wisegreek.com/what-is-a-web-crawler.htm
