ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Survey: Opinion Mining Its Tools and Techniques

Tamanna Sachdeva
Computer Science Department
YMCAUST
Faridabad, India

Divya Kauhik
Computer Science Department
YMCAUST
Faridabad, India

Shruti Sharma
Computer Science Department
YMCAUST
Faridabad, India

Abstract: In previous era the very first step when someone is looking for buying a new product they take advice from relatives, friends or if a company need to launch a new product they perform surveys which was very time consuming task. But today all are busy in their life and have no time for this kind of stuff.

Web being a wonderful substitution for individual opinion asking. Various sites provide different reviews about the product. But size of web is very large. This enormous data again increases the problem of users. So what we need is an automatic technique which provides only valuable information which a user actually requires. Opinion mining is that technique which provides exact opinion from such enormous data about a specific product. It has become the helping hand of today's generation for private decisions from public means.

In this paper we discuss about opinion mining, its types, various tools for accomplishing opinion mining task, and its challenges.

Keywords: opinion mining; polarity; stemming; POS Tagger; morphological analysis; SVM; naïve bayes

I. INTRODUCTION *(HEADING 1)*

Data on web comprises facts and opinions. Current search engines search for facts assuming them true as facts can expressed by using keywords, but they does not extracts exact opinions which are impossible to be extracted using few keywords. Searching for these opinions from web is called opinion mining.

Being rich source for opinions internet is offering a large volume of updated information, the Web data, unfortunately, are typically unstructured text that cannot be directly used for knowledge representation. Moreover, such a huge volume of data cannot be processed manually. Hence, efficient tools and potential techniques are needed to extract and summarize the opinions contained therein.

The formatter will need to create these components, incorporating the applicable criteria that follow.

A. Component of Opinion Mining

- Opinion Holder
- Object
- Opinion

Opinion Holder: Opinion holder is one who is giving the opinion about an object.[4]

Object: An object about which the opinion holder is giving the opinion.

Opinion: Actual sentiment about an object.

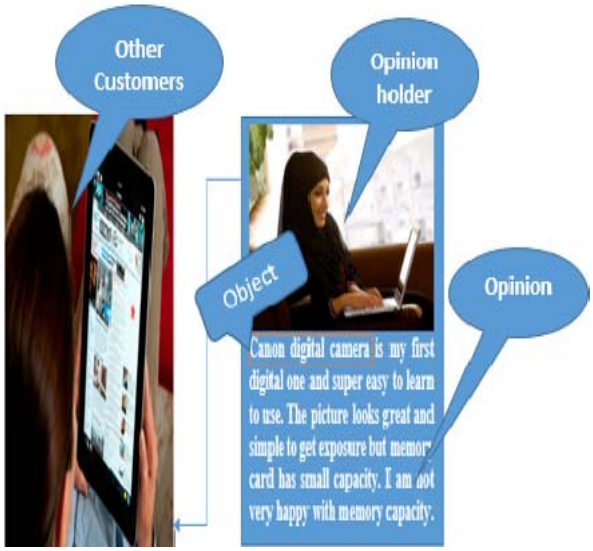

Fig1: Components of Opinion Mining

B. Classification of Opinion MIning

Opinion mining can be classified in three types:

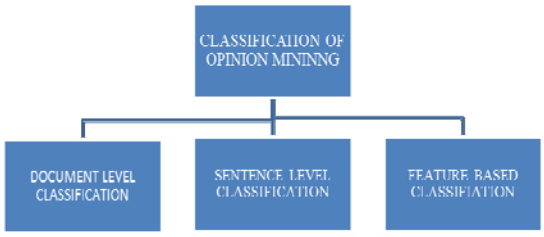

Fig2: Classification of Opinion Mining

1) Document Level Classification: This includes complete document as single entity. One read complete document and then give their opinion according to picture formed in their mind either it is positive or negative.

2) Sentence Level Classification: In this one read complete sentence and tell about the overall polarity of complete sentence.
There are some disadvantages in document and sentence level classification. In those one directly tell either the sentence or document is positive or negative or neutral.

But in normal routine an object is not treated as single entity it always be treated as group of attributes or we can say that group of different features such as a mobile have different number of features such as camera, music player, display etc. and at same time all of us have different choices and we need different features in that particular object such as in mobile someone need better camera and other needs better sound quality. So it is very important to talk about features instead of complete object[6].

3) Feature Based Classification: In this opinion holder talk about the different features of an object. Features might be any attribute of an object or it can be object itself. For example when we are talking about a mobile then it includes various features such as camera, music-player, display, touch etc. In this mobile alone can be treated as an object.

II. ARCHITECTURE OF OPINION MINING

Web data is so vast that it needs to be processed to find the opinionated feature. This opinion gives exact reviews about any object. Various processes are involved in finding that opinion.
An opinion mining architecture [3] contain following components.

A. Pre-Processing

1. The data present on net is so vast that it contains some irrelevant data which is not used for opinion feature extraction. There is a need to remove all those words which have no meaning in finding the opinion feature.

The process which is used to remove those stop words from the reviews obtained from the web is known as pre-processing. The pre-processing contains various steps are involved which helps us to remove those stop words.
1) Morphological Analysis: The first task is to eliminate the irrelevant content such as tags, dates, user name, stop words from the collected review from the web. Then, the noun phrase is extracted from the extracted data which is treated as the candidate feature. For performing morphological analysis, NLP is used which includes POS Tagger.
In NLP, morphological analysis is an essential component which deals with nature of words which are composed of different morphemes.

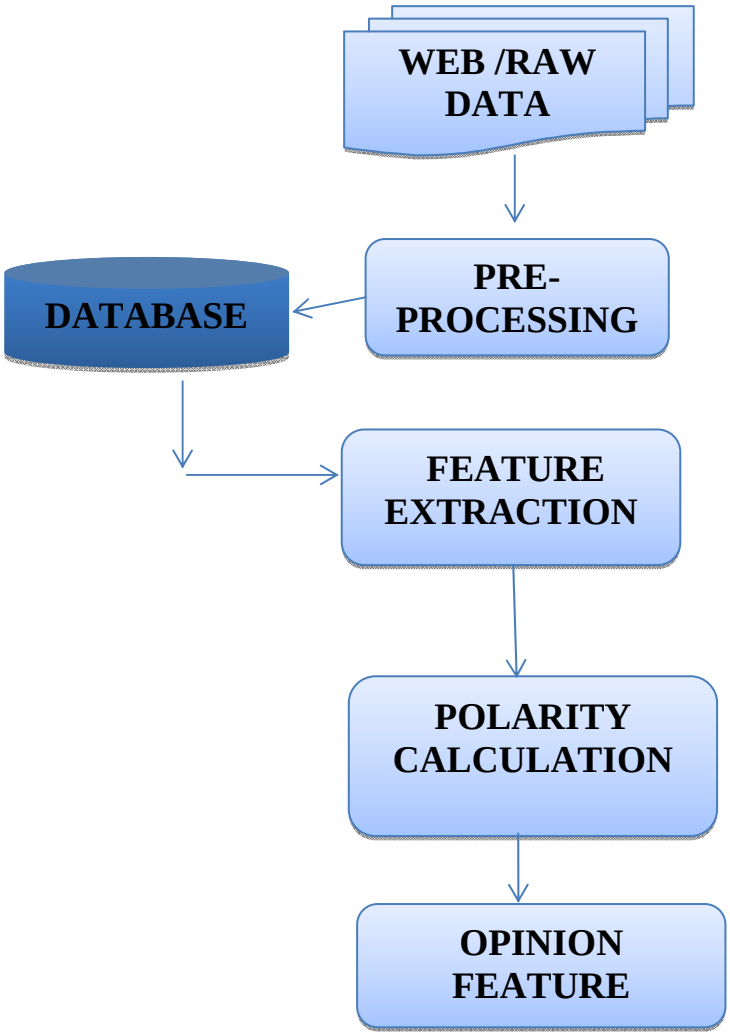

Fig3: Architecture of Opinion Mining

Morphological evaluation recognizes the phrases that the text is made up of and identifies their part of pos.

For example, the end result of the morphological analysis with POS tagging for the sentence “camera quality is good” is as follows:
<NG>*Camera*_NNS</NG> <NG>*quality*_NNS</NG> <VG>*is*_VBP</VG>*good*_JJ.

Here, <NG> and <VG> represent a noun institution (that is, a noun phrase) and a verb organization (that is, a verb word), respectively. Also, POS tagging is done with the aid of assigning and attaching a tag label to each phrase with an underline. Examples of POS tagging labels used consist of nns for a plural noun, vbp for a present annoying verb in non-third character, jj for an adjective. More POS tagging labels can be found in.

After POS tagging, stop-word elimination and stemming are carried out to increase the accuracy of the quest information and the general effectiveness of the device.

Stop-words generally relate to function words, which include determiners (as an instance, ‘the,’ ‘a,’ ‘an’) and prepositions (for example, ‘in,’ ‘on,’ ‘of’), and these stop-words are removed from the sentence on account that they have little

which means on their personal. Stemming is a method of converting variation styles of a phrase right into a common base shape called a stem to reduce the morphological version. For example, 'stemming' and 'stemmer,' are every transformed into the stem 'stem'. Stemming is useful to look for one of these phrases to gain opinion records that incorporate some other phrase inside the equal stem institution.[3]

2) Sentence Splitting: It is the process by which a compound sentence which contains conjunctions is segmented into several simple sentences. The sentence splitting is important because a compound sentence contains various features, each feature represents opinionated information.

Our approach of splitting a compound sentence is carried out by means of spotting an entire clause that is made from a noun phrase and a verb phrase. While numerous whole clauses exist in a sentence, the comma (',') and the connective phrases ('and', 'or', 'however,' and so forth) are used to segregate them.

Hence, the first step of sentence splitting is to divide the enter sentence into numerous candidate entire clauses with the aid of in reality keeping apart the sentence while a conjunctive phrase or the comma is encountered. The subsequent step is to study each candidate clause to see if it is complete, particularly, containing both a noun word and a verb phrase. A candidate clause that meets this condition is identified as a complete clause. On the other hand, a candidate clause that does not fulfill this requirement isn't always whole, and consequently is regarded as an element of the preceding complete clause[7].

B. Feature Extraction

The next stage after pre-processing is feature extraction. In this step, the each sentence which is splitted earlier contains some feature information. The information contains some noun phrases which implies that the feature selection is initiated by the noun phrase.

This noun phrase is considered as a candidate feature. The extraction process of opinion information identifies an adjective which is in form of opinion phrase, if any such phrase is found then it is considered as a proper feature. It also considered negative sentences and replaces the negative words by using negative adjectives phrase.

In feature extraction, feature refinement is also a important task as it reduces the count of features which is obtained by the feature extraction process. Those features which are similar to each other are grouped into single term and all the other redundant features are pruned, this is done with the help of WordNet [10].

C. Polarity

Our next step is to assign the polarity on the basis of extracted and their opinionated information. This polarity may be either positive or negative depending upon the information contained within it.

III. OPINION MINING TOOLS

There are many tools which helps in opinion mining to extract the exact opinion about any object. They work in different stages of opinion mining process. Here we are describing two tools which are used in pre-processing.

A. Stemmer:

Stemmer is a tool which converts variant forms of words into their root words or we can say convert into their stem, and this process of linguistic normalization is known as stemming [12]. Sometime words are converted into valid root word but sometime it gives invalid root word. This work is done under the morphological analysis. Algorithms for stemming have been studied in computer science since the 1960s. Same stem taken as synonyms in many search engines for kind of query expansion, a process called conflation.
For example, 'automatic,' 'automate,' and 'automation' will be converted into the stem 'automat.' Another example which converts words into invalid stem word "argue", "argued", "argues", "arguing", and "argus" reduces to the stem "argu". Whereas "argument" and "argument" will be converted in the stem "argument" which is a valid stem word.
Since starting of stemming there are many stemmers. Let's take a review of them.

1) Lovins stemming algorithm:

Julie Beth Lovins stemmer is the oldest stemmer is the first ever published stemming algorithm [13]. It was remarkable for the early date at which it was done, and for its seminal influence on later work in this area.

The Lovins algorithm is comparatively bigger than the Porter algorithm, as it has very extensive endings list. This lead to big advantage of it: it is faster. With its large suffix set it needs just two major steps to remove a suffix, compared with the eight of the Porter algorithm.

The Lovins stemmer includes 294 *endings*, 29 *conditions* and 35 *transformation rules*. Where its each ending is associated with one of the conditions [14].

Lovins stemming algorithm works in two steps, and these are:

a) First step includes finding of the longest ending which satisfies its associated condition, and is removed.

b) In the second step the 35 rules are applied to transform the ending. This is done whether or not an ending is removed in the first step.

2) Porter's algorithm:

Porter algorithm is given by Martin Porter [15]. It is most commonly used stemmer. It also support java. It works in eight steps. It is also the oldest stemming algorithm by a large margin.

The Porter stemming algorithm (or 'Porter stemmer') is for for removing the morphological and inflexional endings from words in English.
There are some disadvantages in porter's algorithm:

a) There is lack of stemming algorithms for languages other than English.

b) It is very similar to "porter stemmer" but with slightly improved rules.

c) It is very slow algorithm

3) Porter's 2 algorithm:

Martin Porter had developed a language to write stemming algorithm [15]. This is the successor of Porter's algorithm developed by Martin Porter. This is faster than porter's algorithm. It has more improved rules than its predecessor. Previously all the stemming algorithms were developed for English language only. This is the first multilingual stemming algorithm

B. POS Tagger

POS-tagger is a tool that gives the part of speech used in the sentence. Need of POS Tagging is to extract features or opinion information from the sentence. After applying POS-Tagger on a sentence or document it gives assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

In schools we are taught 9 part of speech these are noun, verb, adjective, preposition, proverb, adverb, conjunction and interjection. However, there are clearly many more categories and sub-categories.

Whereas in part-of-speech tagging by computer, it gives more perfect way to distinguish from 50 to 150 separate parts of speech for English. For example, NNS for plural common nouns, NP for singular proper nouns, NN for singular common nouns

IV. TECHNIQUES OF OPINION MINING

In opinion mining major tasks includes feature selection, classification, optimization etc. For these tasks we can divide the techniques in following major categories—

1. Supervised Learning→ In supervised learning we have some sort of training data and some test data, by using training data it make some function which is tested by test data. After checking it can be used for new examples for classification. Here we can say input is attached with associated output[8].
2. Unsupervised Learning→ In unsupervised learning no training data is given to the learner, in this no output is attached with input, everything is done with help of observations. One of the main technique in unsupervised learning is clustering, a technique in which objects having similar characteristics are grouped together.

Table 1

Sr. No.	*Author Name*	*Title*	*Technique Used*	*Advantage and Disadvantage*
1.	Bo Pang, L Lee S Vaithyanathan	Thumbs Up? Sentiment Classification Using Machine Learning Techniques	Naïve Bayes SVM	**Advantages of Naïve Bayes** Simple in use Optimal for highly dependent features [2] **Disadvantage of Naïve Bayes** Because of taking features independent to each other, it does not hold good in real world. **Advantages of SVM** High accuracy Good when your data is linearly inseparable.
2	Alessia D'Andrea Fernando Ferri Tiziana Guzzo Patrizia Grifoni	Approaches, Tools and Applications for Sentiment Analysis Implementation [9]	**Machine Learning (**Bayesian Networks Naive Bayes Classification Maximum Entropy Neural Networks SVM **)** **Lexicon based** (Dictionary based approach Novel Machine Learning Approach)	**ADVANTAGES** Ability to adapt and create trained models for specific purposes and contexts **LIMITATIONS** Low applicability to new data because it is necessary the availability of labelled data that could be costly or even prohibitive **ADVANTAGES** Wider term coverage **LIMITATIONS** Finite number of words in the lexicons and the assignation of a fixed sentiment orientation and score to words
3	S. Kasthuri, Dr. L. Jayasimman, Dr. A. Nisha Jebaseeli	An Opinion Mining and Sentiment Analysis Techniques:A Survey	Naïve Bayes And K Nearest Neighbour[11]	**Advantages of Naïve Bayes-** Easy to implement and good even when applied to large database. It requires small set of practical data.[1] **Advantages of KNN-** It is simple even in solving complex problems. This is also called lazy learner. Effectively uses hypothesis space. Training is fast **Disadvantages of KNN-** More time taken in prediction.
4	Zhongwu Zhai, Bing Liu, Hua Xu, Peifia Jias	Clustering Product Features for Opinion Mining	Expectation-Maximization(EM) algorithm	**Advantages of EM-** EM resolves the problem of semantically same words for features.[6]

V. CHALLENGES OF OPINION MINING

Various challenges in opinion mining are-

1) Feature Extraction:

Feature extraction is the main task in feature based classification. But this task includes various problems. [5] One of the main problem is various works in this fields consider only noun words as features, whereas non-noun words can be the features. For example in case of mobile display and touch are non-noun words but are the important features of the mobile.

2) Polarity Calculation:

Polarity calculation means find out either given review is positive or negative or neutral. In some environment the review act as positive and in other it behaves as negative. So it is hard to decide actual polarity.[3]

3) Different behaviour of people:

Sometime there is a case for some person the review is positive and for other it is negative.[7] Another problem in this is different people write comments in different languages. Even in using same languages there way of writing is different.

4) Irregularity in defining Neutral opinion:

This is the least worked area in opinion mining. What should be the various parameters in this analysis is hard to decide.

5) Document and sentence level classification:

Document and sentence level classification has a major disadvantage as in this either complete document or a sentence is taken as whole and according to that it gives polarity decision either it is positive or negative. But in document or in a sentence it is not always possible that it contain either positive or negative review. Sentence may be compound, it means that may have both positive and negative opinion at the same time.

VI. REFERENCES

[1] Julie Beth Lovins, Development of a Stemming Algorithm [Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, March and June 1968] * by, Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

[2] Bo Pang and Lillian Lee Shivakumar Vaithyanathan Thumbs up? Sentiment Classification using Machine Learning Techniques Department of Computer Science Cornell University arXiv:cs/0205070v1 [cs.CL] 28 May 2002

[3] Hana Jeong, Dongwook Shin, and Joongmin Choi FEROM: Feature Extraction and Refinement for Opinion Mining ETRI Journal, Volume 33, Number 5, October 2011

[4] Bing Liu Hua Xu Peifa Jia Zhongwu Zhai -Clustering Product Features for Opinion Mining WSDM'11 , February –12, 2011, Hong Kong, China. Copyright 2011 ACM

[5] Bakhtawar Seerat, Farouque Azam Opinion Mining: Issues and Challenges (A survey) International Journal of Computer Applications (0975 –8887) Volume 49–No.9, July 2012 42

[6] Nidhi Mishra C.K.Jha Classification of Opinion Mining Techniques International Journal of Computer Applications (0975 –8887) Volume 56–No.13, October 2012 1

[7] Raisa Varghese, Jayasree.M- A SURVEY ON SENTIMENT ANALYSIS AND OPINION MINING IJRET: International Journal of Research in Engineering and Technology Volume: 02 Issue: 11 | Nov-2013, Available @ http://www.ijret.org 312

[8] Asmita Dhokrat Sunil Khillare C. Namrata Mahender Review on Techniques and Tools used for Opinion Mining International Journal of Computer Applications Technology and Research Volume 4 – Issue 6 , 419 -424, 2015, ISSN:-2319–8656www.ijcat.com 419

[9] Alessi D Andrea, Fernando Ferri Patrizia Grifoni, Tiziana Guzzo International Journal of Computer Applications (0975 –8887)Volume 125 – No.3, September 2015 26 Approaches, Tools and Applications for Sentiment Analysis Implementation

[10] Package'wordnet' January 6, 2016 Title WordNet Interface Version 0.1-11

[11] S. Kasthuri Dr. L. Jayasimman Dr. A. Nisha Jebaseeli An Opinion Mining and Sentiment Analysis Techniques: A Survey International Research Journal of Engineering and Technology(IRJET) Volume: 03 Issue: 02 | Feb-2016

[12] https://en.wikipedia.org/wiki/Stemming

[13] http://snowball.tartarus.org/algorithms/lovins/stemmer.html**.** The Lovins stemming algorithm

[14] http://snowball.tartarus.org/algorithms/porter/stemmer.html The Porter stemming algorithm

[15] http://snowball.tartarus.org/texts/introduction.html.Snowball:A language for stemming algorithms.