Fine Grained Opinion Mining Techniques

Lakshmisampanna N¹, Prof. Kavita P. Shirsat²

¹ Department of Computer Engineering, VIT Wadala, Mumbai University, India.
² Professor, Department of Computer Engineering, VIT Wadala, Mumbai University, India.

Abstract—Today, the world is moving towards web due to its growth in Internet. Internet is playing a vital role in people’s life. From playing games online to buying products, internet is serving all purposes. Before buying a product, people see the product’s feedback or opinions. Opinions play a vital role due to its capacity to change the buyer’s decision for a product. Opinion mining is a growing area where these opinions are mined to extract knowledge. This paper mainly does a review of different opinion mining techniques and tries to suggest our take on it.

Keywords—Opinion mining, Aspect level mining, Reviews, Polarity, Aspect extraction, Opinion orientation.

I. INTRODUCTION

Data mining these days has grown tremendously which allows researcher or the data miner to perform different experiments and has tremendous opportunity. Opinion mining being an important field in data mining helps in identifying the user opinion about the respective feature. Online shopping is done extensively these days and people before doing any purchase have a tendency to go through the reviews about the product.

The opinions present in the websites are huge in quantity and it is very difficult to read all of them. So mining of these opinions helps in summarizing as well as bringing out the required knowledge.

There is a lot of scope and challenge in this opinion mining area which gives the researchers an opportunity to play with the data.

A. Opinion mining

Opinion mining is a specialized area coming under Data mining, specifically under web content mining. Opinion mining is defined as extracting people’s opinion from the web. It analyses people’s opinions, appraisals, attitudes and emotions towards organizations, entities, persons, issues, actions, topics and their attributes as in [1].

The goal of Opinion mining is to make system able to recognize and express emotions. A thought, view, or attitude based on emotion instead of reason is called sentiment.

Users expresses their view about the products that they have bought and experienced using blog posts, shopping sites, review forums etc. Sentimental analysis or Opinion mining aims to extract these opinions and tries to draw knowledge from it.

B. Levels of Opinion mining

Opinion mining is a method of tracking the feelings of a person about a particular item, company, events and issues. It analyzes which part has opinion expressing, who wrote the opinion and what is being commented in online reviews [2].

The general categorizations of Opinion mining is as follows [1],

- Document-Level
- Sentence-level
- Aspect/Phrase-level

1) Document – level opinion mining:

The basic information unit is the whole document. At this level the whole document is classified as positive or negative. Sometimes this is not that much preferred because in the document there can be many aspects which may not be identified. This means if polarity of the document is positive, it doesn’t mean that document has positive opinion for each aspect. Similarly if polarity of the document is negative, it doesn’t mean that for each aspect in the document the polarity is negative.

2) Sentence – level Opinion mining

Similar to Document level Opinion mining, sentence level opinion mining categorizes sentences instead of the whole document. Here the document is broken down into individual sentences and Opinion mining is applied to each sentence. Here the outcome will be either the sentence is positive, negative or neutral. Similar to document level there is an uncertainty whether the sentence classified as positive/negative has all its individual aspect positive/negative respectively. But compared to document level, sentence level has better clarity.

3) Aspect/phrase- level Opinion mining

This is the fine-grained mining level where individual sentences are broken down with respect to aspects and their polarity is found. Aspects are the features present in any user comment with respect to which he/she has an opinion to make on. To determine every opinion of the document, aspect level mining is mostly preferred. Aspects found in the user opinion can be further classified into Explicit Aspect and Implicit Aspects as in [3].

- Explicit Aspects: Features that are directly specified and can be easily identified in the reviews, these are noun and noun phrases.
  E.g.: “The picture quality of this phone is great”. Aspect here is directly specified i.e. “Picture Quality”.
Implicit Aspects: Features that are not directly specified nor can be easily identified, these are not noun or noun phrases. E.g.: “This phone doesn’t fit into my pocket”. Aspect here is “Size” of the phone which is not directly specified.

II. RELATED WORK
The automatic analysis of Opinion Mining has opened up many doors to experiment with the data and the knowledge that is obtained from it is also used extensively nowadays.

A Jeyapriya et al [2] proposed an opinion mining technique which was in the aspect level. It used supervised learning algorithm to classify the opinions based on their orientation.

Richa Sharma et al [4] went through a dictionary based approach of the unsupervised technique to determine the orientation of the sentences. Here the polarity of the given sentence is calculated on the basis of majority of the opinion words.

Mita K. Dalal et al [5] underwent a semi-supervised learning methodology based opinion summarization and classification of online products reviews. Feature pruning is also performed by using decomposition rule for multiwords, manual inspection etc.

Mugdha More et al [6] proposed robust feature-based opinion summarization system based on weighting scheme and association rule. Ensemble technique is used for feature extraction and orientation.

III. METHODOLOGY

As shown in the Fig.1 the general schema of the Opinion mining system consists of a review database which is the collection of the reviews from general or targeted websites. To this review collection, pre-processing steps are performed.

1) Pre-processing
Pre-processing step is very important for cleaning the data which in turn improves the whole mining process. Some of the commonly used pre-processing steps are
- Stop word removal
- POS Tagging
- Stemming etc.

Stop word removal: Stop words are language specific functional words which carry no information. It may be of types such as pronouns, prepositions, conjunctions. Stop word removal is used to remove unwanted words in each review sentence. Words like is, are, was etc.

POS (Parts Of Speech) Tagging: POS tagging is the task of labelling (or tagging) each word in a sentence with its appropriate part of speech. POS tagging is an important phase of opinion mining, it is essential to determine the features and opinion words from the reviews.

Stemming: It is the process for reducing inflected or sometimes derived words to their stem or root form.

2) Aspect Extraction
Once the initial pre-processing work is done then the reviews or the pre-processed data is ready for further processing i.e. Aspect extraction. Aspect extraction is defined as selecting aspects from the reviews. Here aspects are generally the objects of which the opinion is expressed. Whenever the POS is conducted, it identifies the noun or noun phrases which generally are the aspects. Consider an example.

“The voice quality of this phone is extremely good". This example illustrates that “voice quality” is the object about which the opinion is given. So Aspect is “voice quality”.

There can be multiple aspects in any review and multiple opinions with respect to each aspect. So mining these reviews gets complicated as its gets deep into the analysis process.

According to the paper [2], they have used the concept of frequency count. This mainly uses frequent item set mining which adopts minimum support count concept. There is a threshold set and noun or noun phrases which exceed the count of the threshold will be stored. Here the algorithm first extracts the entire noun and noun phrases in each reviews and stores it in a text file. Minimum support count is used to find all frequent aspects from review sentences. Aspects like battery, voice quality, memory etc. Then frequent aspects are extracted and stored [2].

The paper [7] divides the entity extraction into two i.e. product aspect extraction and opinion extraction. Where the aspect extraction is done based
on frequent sets i.e. analysing the purchase behaviour of the buyer [7]. The main idea here is to search frequent “tags” sets which can be understood by analysing how people tend to express their feelings in normal language. Here initially to get maximum result, first all the product specifications are collected and stored and then they are expanded with the help of list by word synonyms [7]. Following list shows the illustration of frequent tags as in [7].

- [NN][VBZ][RB][JJ] e.g. :“Software is absolutely terrible”
- [NNS][VBP][JJ] e.g.:“Pictures are razor sharp”
- [NN][VBZ][RB][JJ] e.g.:”ear piece is very comfortable”

Where NN - Noun, singular or mass
NNS - Singular/Plural Noun
VBZ - Verb, 3rd person singular present
VBP - Verb, non-3rd person singular/Plural
RB - Adverb
JJ - Adjective

These are some POS tags.

In the paper [4] they do not give much importance to the aspect extraction as compared to the detailed opinion extraction and orientation. Here Simple POS tagger is used to tag all the reviews and then aspects are extracted as they are mostly noun or noun phrases.

In paper [5] they have also used frequently occurring noun or noun phrases as possible opinion features. They perform POS tagging using Link Grammar Parser. Here they generate a frequent feature set using the multiword approach.

3) Opinion orientation

Once the aspects are extracted their relative opinion should be classified as positive or negative. This process of orientation of user opinions into their respective class is called Opinion orientation. This process is very tricky and difficult because understanding user reviews automatically is not that easy. There can be some indirect meaning to some statements or some sarcastic sentence which is very hard to crack. It certainly requires some natural language processing techniques which to some extent can decode it but not completely.

Paper [2] suggests some interesting steps for sentence and aspect orientation. They use the concept where initially they try to identify number of positive and negative sentence in reviews. This is done by identifying the opinion words. Positive and negative labels are considered to be collective opinion words. Good, bad, excellent, not good etc are some examples of opinion words. Next step is identifying the number of positive and negative opinions for each collected aspect. Here Naïve Bayesian algorithm using supervised term count approach is used. Both sentence and aspect orientation is implemented using Naïve Bayes classifier. 92.37% accuracy is obtained using this approach.

The Paper [7] does extraction of opinion once aspects are extracted. They propose that usually an adjective represents the opinion and they will be present nearby aspects in the sentence. They first locate adjectives or the opinion words and then they start searching backwards for the nearest aspect. If aspect not found then search will be opposite i.e. forward search is conducted.

They use Opinion lexicon developed by Hi and Liu [8,9] to extract opinion words. If word in the sentence matches positive dictionary then the polarity of the opinion word is positive else negative.

Paper [4] uses an unsupervised technique where after extracting the aspect, opinion extraction and seed list preparation is done. Initially it uses dictionary approach where some of the common opinion words along with their polarity are stored in the seed list. All the opinion words are extracted from POS tagged output. Here words are matched with seed list. If match found, polarity is found in the seed list. If match not found, then its synonyms are checked from the WordNet dictionary and if match found then same polarity from the seed list is stored and seed list is updated. If not, antonym is checked and if a match is found then extract and store the opinion word along with opposite polarity and seed list is updated. Else, process is repeated.

In Paper [5] while extracting the aspects they even store the describing adjectives as well as modifiers if present. And in order to describe the sentiment polarity of the adjective describing the aspect SentiWordNet, a lexical resource for opinion mining is used. It assigns three normalized sentiment score-positivity, negativity and objectivity. If positive score is high for any word then it is classified as positive and if negative score is high then word is classified negative. And hence the opinions are oriented.

Above are some strategies analyzed by us while dealing with mining techniques.
IV. PARAMETERS FOR EVALUATION

The performance of the system is evaluated. Precision, recall and F-measure are the parameters used in the system for evaluation.

Precision is the measure of retrieved instances that are relevant. Recall is the fraction of relevant instances that are retrieved. F-measure is a measure of test’s accuracy as in [2].

Precision, recall and F-measure are defined as follows as in [2],

Precision\(=\frac{|\text{Extracted values} \cap \text{True values}|}{|\text{Extracted values}|}\)

Recall\(=\frac{|\text{Extracted values} \cap \text{True values}|}{|\text{True values}|}\)

\[F\text{-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}\]

To calculate these measures, true values in reviews are identified manually. Then system mines the aspects and opinions (extracted values). Using this, precision, recall and F-measure are calculated for customer reviews.

V. CONCLUSION AND FUTURE WORK

This analysis is done mainly due to the growing interest in the field of Opinion mining. Nowadays every buyer before buying the products irrespective of online purchase or offline purchase has a habit of going through these reviews. Mining these reviews will be a boon to everybody whether it is a manufacturer or buyer.

Whenever reviews are concerned, trust factor comes in to the picture. Can we believe the reviews? Is it trustworthy? Because these days there are spam reviewers who mislead customer by writing untruthful reviews.

In future work we are planning to build a system which first tests the trustworthiness of the reviews and the reviewer. And if the review and the reviewer is proved to be trustworthy only then his/her review is considered for the opinion mining process. Hence if trust factor is induced to opinion mining we believe there will be a certain degree of trust that will be built. This in turn helps opinion mining process with respect to its result.

REFERENCES


