Feature Extraction From Product Review Using Ontology

Submitted By

Drashti Naik

15MCEI16

DEPARTMENT OF COMPUTER ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2017

Feature Extraction From Product Review Using Ontology

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

(Information and Network Security)

Submitted By

Drashti Naik

(15MCEI16)

Guided By

Prof.Jitali Patel

DEPARTMENT OF COMPUTER ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2017

Certificate

This is to certify that the major project entitled **"Feature Extraction From Product Review Using Ontology"** submitted by **Drashti Naik (Roll No: 15MCEI16)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Jitali Patel
Guide & Assistant Professor,
CE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sharada Valiveti
Associate Professor,
Coordinator M.Tech - CSE (INS)
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Alka Mahajan
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Drashti Naik**, Roll. No. **15MCEI16**, give undertaking that the Major Project entitled "**Feature Extraction From Product Review Using Ontology**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering (INS)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Drashti Naik(15MCEI16)

Date:

Place:Nirma University,Ahmedabad

Endorsed by

Prof. Jitali Patel

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Jitali Patel**, Associate Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for her valuable guidance and continual encouragement throughout this work. The appreciation and continual support she has imparted has been a great motivation to me in reaching a higher goal. Her guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr.Alka Mahajan**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation she has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Drashti Naik

15MCEI16

Abstract

Opinion mining is accepting more attention because of the development of blogs, e-commerce, news, reports, forums and additional web sources where individuals tend to express their opinions. Different people have different opinions. Peoples thought may vary according to the domain and opinion may contain both positive and negative words. For a product, user may like or dislike some of its features. Filtering this review and extract domain related features is the important task of this paper. In this paper, ontology is used to extract the features and adjectives are used as the sentiment word. Sentiment Analysis is used to obtain positive or negative feature of the review.

Abbreviations

POS	Part Of Speech
NLTK	Natural Language ToolKit
LDA	Latent Dirichlet Allocation
API	Application program interface
SA	Sentiment Analysis
ABSA	Aspect Based Sentiment Analysis

Contents

List of Figures

Chapter 1

Introduction

With the expanded utilization of the Internet around the world, the development of online business is very high, and purchasing items from the online websites has turned into another pattern in the present culture. There are a large no of items accessible in the market and its expanding day by day. Now, everything is online, people can also share their opinion online and make decisions. More and more people post reviews for any product, movies, event and political thought. From all of this review finding proper opinion is time consuming and burdensome. It is common for a person to learn what others like and dislike before buying anything and for a manufacturer to keep track of their opinion on its product to improve the satisfaction of the customer and provide them good services. So, now a days there are so many research available for sentiment analysis and recommendation system. Through the sentiment analysis customer can easily get the opinion for particular product. From research perspective, a product features can be express in the form of noun, adjective,verb or adverb. From this extract the feature and apply any machine learning algorithm for predict positive and negative review.

For measuring the quality of various items and services online reviews are essential. With the quick development in the number of electronic documents and a large amount of data on the Web, people, business substances, and different associations are looking for the better approach to extract and use public opinion. Large datasets are accessible online, they can be the numerical or content record. The amount of data accessible on the web is vast and the development of methods for automatic categorization and organization of this information have been the focus of many researchers.

For the last few years, social media have extended exponentially with so many individuals share their opinions. This freely accessible data can be utilized to produce value, such as predicting the electronics sales, stock market, and box office collection. However, the use of sentiment analysis is not restricted to predict the behavior but can also be used to examine the standing of a company, product or any other entity with regards to public opinion.

If companies could identify the aspects of their products, which are frequently discussed in a negative way online by the customer, they could better check them and improve customer satisfaction. Moreover, if companies could identify which aspects of their products customers like that data can be used to develop those features into other products or use them for marketing and product development purposes.

There is a lot of information on the Web. Textual information can be divide into two categories, they are facts and opinions. Facts are the objective statements and there is a lot research for the information extraction. However, opinions are the subjective statements and still need the further research for getting the information of the peoples opinions or sentiments in their mind.

People who buy or sell the products give their comments, feedback, etc in the form of text which is mostly unstructured. It becomes necessary to categorize these texts to make business intelligent solutions. The opinion mining on blogs is the essential and trivial task now a days to explore the product features. The huge data available in the internet has to be modeled, analyzed and then the decision has to be taken. Retrieving the information from the unstructured text is the difficult task. The evolving semantic web technologies can be used to overcome these problems to represent domain vocabularies and their relationships through ontologies, RDF, etc to make this task easier.

The main objective of this work is to extract more and more features related to domain and help producer and consumer.

1.1 Sentiment Analysis:An Overview

SA, also called opinion mining, is the field of study that analyzes peoples opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products,

services, organizations, individuals, issues, events, topics, and their attributes.[1] Its aim is to detect subjective information from any feedback and review given by the people. Opinions are the key influence of people's behaviors as they are important to any human activity online.

Sentiment analysis is different from the traditional text mining process. The traditional text mining focuses on the analysis of facts from unstructured text, whereas sentiment analysis deals with attitudes of people. The issue is to locate the relevant information and to get a general opinion of what individuals think about an entity. Essentially using the ratings given is not sufficient to get a description of what individuals think. Sentiment analysis is associated to not only with opinions but also with the emotions, feelings, and attitudes of individuals.

Opinions from people are the key influences of their behaviors. Beliefs and perceptions of reality are conditioned on how others see the world. Whenever the decision has to be made, people often ask opinions from others. Organizations use surveys, focus groups, opinion polls, consultants and the like to get different opinions. People also express their views through email, blogs, social networking sites like Facebook, Twitter, etc.

SA alludes to the utilization of NLP and text analysis to distinguish and separate subjective data from source data. There are so many application for which sentiment analysis is connected to social media and reviews.

Three levels for SA:

Document Level:

It classifies whether the whole document expresses positive or negative opinion. It assumes that for a single product there is a document which expresses the opinion. So, when there is a comparison between two product this analysis is not applicable.The input can be reviews, blog, articles, tweets etc.

Sentence Level:

Its check the sentence one by one from the document and express positive, neutral or negative opinion. The issue with document level or sentence level is that, we can't get specific information, for example, positive or negative opinion in regards to particular product features from the reviews.

"I bought the mobile last week. The picture quality is very impressive. The battery life is not that great. It is worth the price." In document level opinion mining, the above review may be classified as positive. But the reviewer discusses different features of the phone and expresses different opinions about each. The whole data is not captured by document level or sentence level opinion mining techniques. To a consumer, these detailed information may be necessary for decision making. The aspect level opinion mining tries to address this issue.

Entity and Aspect Level:

It is also called feature level sentiment analysis. It checks for sentiment, whether it is negative or positive and target for which the opinion is given. For example, "Camera is not good,but i still like the phone" has a positive tone, but we can't say it's positive for the whole sentence.

Every review consist of opinion targets, sentiments, opinion holder and time, which is useful for predicting any review. Opinion target is for which product the opinion is given, Sentiments is the opinion given for the product is positive or negative, Opinion holder is the one who give the review of the product and Time is when the opinion holder given the review.

Types of Opinion:

Two types of opinion regular and comparative opinion. Regular opinion is divided into two parts Direct and Indirect.

Direct Opinion:

It expresses opinion directly on an entity and entity aspect. "JBL headphone sound quality is great." express a positive opinion.

Indirect Opinion:

It expresses opinion indirectly on an entity and entity aspect which effects other entity. "After eating this dog food, my dog weight is decreased." express effect of food on dog health and give negative opinion.

Comparative opinion:

It expresses the difference and similarity between the two or more entity and entity aspect. "iPhone is better than Blackberry"

There are basically two types of review user can give.One format is user write the review in the form of advantages and disadvantages. Another format is user write the review in the form of full-text format.[1]

The sentiment classification techniques are generally divided into two categories:

- Machine Learning Approach
- Lexicon Based Approach

A third kind of classification is hybrid approach, which utilizes both machine learning techniques and lexicon-based approach. The different methods and techniques in the above two approaches are represented in Figure 1.1.

Machine Learning Approach:

Different machine learning algorithms are used in this method. During the training process, Machine learning algorithms basically learn and store qualities of the information. For the testing data best category data are determine from the knowledge. The training and testing data are explained by labels. On the basis of data size different validation techniques can be applied. For SA evaluations cross-validation is used. The dataset is divided into i parts, then initial segment is consider as the testing data and the other as training data.

Lexicon Based Approach:

This approach depends on the intuition that the polarity of a piece of text can be acquired on the ground of the polarity of the words which compose it. For the content analysis opinion lexicon is used. So the efficiency of the technique relies on the integrity of the lexicon resource.

In SA, feature extraction is one of the most complex tasks, since it requires the use of Natural Language Processing techniques in order to automatically identify the features in the opinions under analysis.

1.1.1 Challenges Of SA

There are some challenges while filtering the reviews for the sentiment analysis.

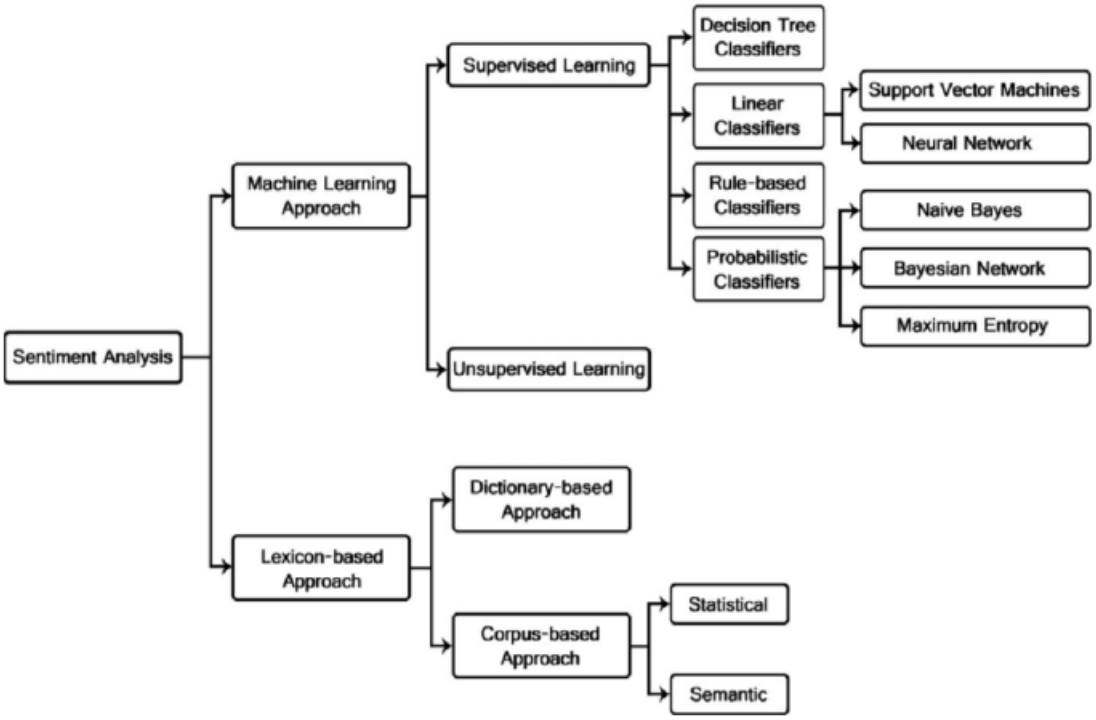

Figure 1.1: Sentiment Classification Techniques

For the various application domains positive and negative opinion word may have inverse orientation. "loud" is negative word but for some sentences it can be positive like "the music is very loud its perfect for the party".

Sometimes sentence having a opinion word can't explain any sentiment. This occurs in conditional sentence and question, e.g. "Which phone is perfect for the camera?" Sentence have a positive word, but it is not a positive or negative sentiment for a phone.

In some cases, a sentence without sentiment word may also express some opinion. This sentences generally consider some precise knowledge in an objective way. "This phone uses a countless battery" have a negative opinion about the phone since it uses more resources.

There are a few circumstances in which a more complicated definition is required. For example, "This door length is small for a tall person," which don't state that the door is too small for all. The context of the sentiment is an necessary information.

Blogs are maybe the tough to manage, as people can argue and interact with each

other on any topic. Various application domains are additionally viewed as extremely hard to manage. Social and political talks are significantly tough than sentiments for the products and services, because of complicated subject and opinion expressions. On the social site, someone can also write the opinion in the sarcasm way.

There are people who gave bogus opinions to elevate or to dishonor other organization's products or services. They are called opinion-spammers and the forged opinions are called opinion spam.

1.2 Ontology:An Overview

An ontology is a formal description of concepts in a domain of discourse (classes), properties of every concept describing various features and attributes of the concept, and restrictions on attributes. In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.

Ontology is generally considered as a formal specification of conceptualization which consists of concepts and their relationships. Domain ontology is one kind of ontology which is used to represent the knowledge for a specific type of application domain. Ontologies provides a formal, structured knowledge representation, with the advantage of being reusable. They also provide a common vocabulary for a domain.

OWL is a stable specification developed by the Web Ontology Working Group. It is considered a Web standard for industry and academy. Ontology can reflect the nature of the objective things and its external performance and the field of artificial intelligence, ontology can help us to get the essential knowledge of things.

Semantic Web technologies are currently achieving a certain degree of maturity. The Semantic Web was conceived with the aim of adding semantics to the data published on the Web thus allowing machines to be able to process these data in a way similar to that of humans. Semantic Web technology may be a valuable addition to traditional opinion mining approaches. More concretely, ontologies constitute the standard knowledge representation mechanism for the Semantic Web and can be used to structure information. The formal semantics underlying ontology languages enables the automatic processing

of the information in ontologies and allows the use of semantic reasoners to infer new knowledge.

For ontology creation there are some steps.

- Determine the domain and scope of the ontology

  Consider the ontology of cellphone. Representation of cellphone feature is the domain of the ontology. We can use this ontology for the applications that suggest good features of the phone based on the customer requirement. So, this is it's scope.

  The concepts describing different types of features, main types, the notion of a good and a bad feature will figure into the ontology. If the ontology we are designing will be used to assist in natural-language processing, it may be important to include synonyms and part-of-speech information for concepts in the ontology. If the ontology will be used to help customers decide which phone to buy, we need to include pricing information.

- Consider reusing existing ontologies

  It is almost always worth considering what someone else has done and checking if we can refine and extend existing sources for our particular domain. Reusing existing ontologies may be a requirement if our system needs to interact with other applications that have already committed to particular ontologies.

  Many ontologies are already available in electronic form and can be imported into an ontology development environment. There are libraries of reusable ontologies on the Web. For example, we can use the Ontolingua ontology library (http://www.ksl.stanford.edu/software/ontolingua/) or the DAML ontology library (http://www.daml.org/ontologies/).

- Define the classes and the class hierarchy

  A top-down development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts. For example, first create classes for the general concepts of phone. Then specialize the phone class by creating some of its subclasses like processor, storage, etc. Further categorize the processor into single core, dual core and so on.

A bottom-up development process starts with the definition of the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general concepts. For example, start by defining classes for Email and SMS. Then create a common superclass for these two classes name communication.

A combination development process is a combination of the top-down and bottom-up approaches. Define the more salient concepts first and then generalize and specialize them appropriately.

Whichever approach is used, usually classes are define first. Select the terms that describe objects having independent existence rather than terms that describe these objects. These terms will be classes in the ontology and will become anchors in the class hierarchy. Organize the classes into a hierarchical taxonomy. If a class A is a superclass of class B, then every instance of B is also an instance of A.

- Define the properties of classes

  Once classes are defined then the internal structure of concepts is defined. The remaining terms are likely to be properties of these classes. These terms include, for example, a phones color, body, size, weight, etc. For each property in the list, we must determine which class it describes.[16]

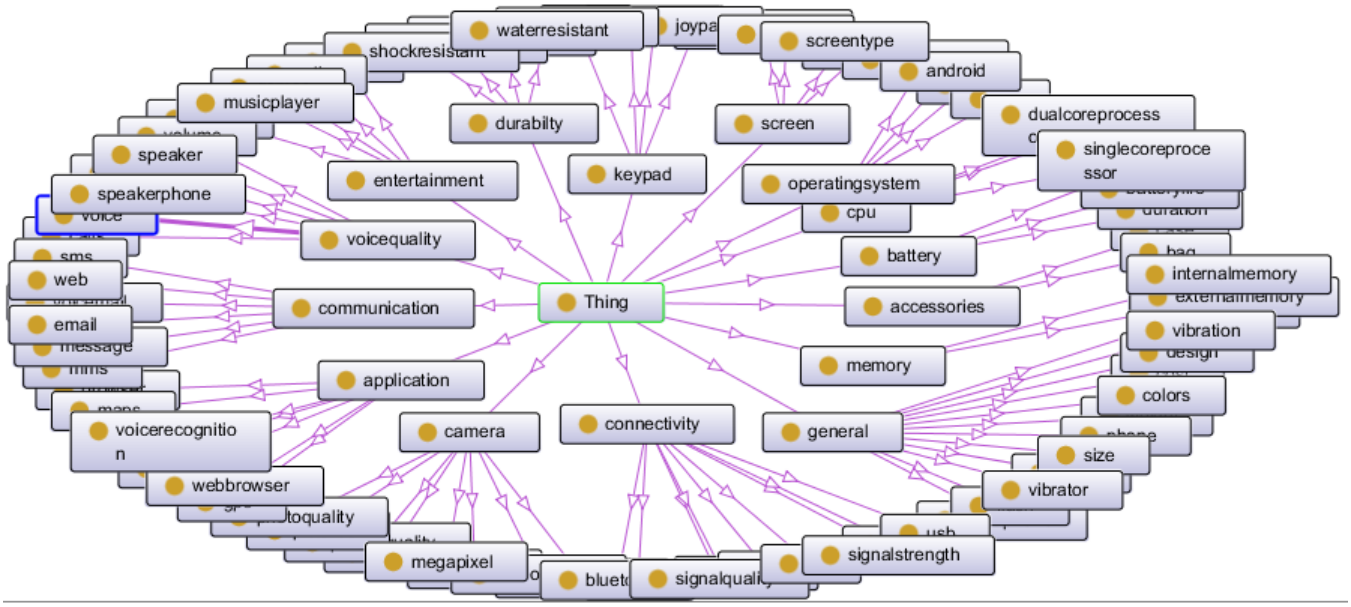

Figure 1.2: Graphical view of CellPhone Ontology

Components of Ontology:

Classes:

Classes are also called as sets, collections, concepts, or types of objects. Classes are generally the overview of the ontology. A class can subsume or be subsumed by other classes. A class subsumed by another is called a subclass.

Attributes:

Attributes are also called aspects, properties, features, characteristics, or parameters that objects (and classes) can have. Objects in an ontology can be described by relating them to other things, typically aspects or parts. These related things are often called attributes. Each attribute can be a class or an individual. The kind of object and the kind of attribute determine the relation between them. A relation between an object and an attribute express a fact that is specific to the object to which it is related. For example, <has as name> iPhone.

Relationships:

Relationships (also known as relations) between objects in an ontology specify how objects are related to other objects. Typically a relation is of a particular type (or class) that specifies in what sense the object is related to the other object in the ontology. For example in the cellphone ontology that contains the concept Single Core and the concept Dual Core might be related by a relation of type ¡is defined as a successor of¿. The full expression of that fact then becomes: "Dual Core is defined as a successor of : Single Core" This tells us that the Dual core is the processor that replaced Single Core.

Restrictions:

Formally stated descriptions of what must be true in order for some assertion to be accepted as input.

Rules:

Statements in the form of an if-then sentence that describe the logical inferences that can be drawn from an assertion in a particular form.

Function terms:

Complex structures formed from certain relations that can be used in place of an individual term in a statement.

Events:

The changing of attributes or relations.

For ontology creation there are so many tools like Protege, Java Ontology Editor, Ontolingua, Chimaera, OilEd, etc. Generally we can save the ontology in the form of OWL, RDF or XML format. There are some Languages, which support ontologies management.

ACRONYM	LANGUAGE	CHARACTERISTICS
HTML	Hyper Text Markup Language	Simplicity
XML	Extensible	Extensions for arbitrary domains and specific tasks.
SHOE	Simple HTML Ontology Extensions	It is a XML compatible knowledge representation language for the web. It allows page authors to annotate their web documents. It is not actively maintained.
RDF	Resource Description Framework	Syntactic conventions and simple data models to represent semantics. It supports interoperability aspects with object–attribute–value relationships.
RDFS	Resource Description Framework Schema	Primitives to model basic ontologies with RDF.
OIL	Ontology Inference Layer / Ontology Interchange Language	Primitives to model ontologies from frame–based languages, formal semantics and reasoning support based on descriptive logic, a proposal for syntatic interchange of annotations. It is compatible with RDF Schema [Fensel et al. 2000].
DAML	DARPA Agent Markup Language	It is formed by DAML–ONT, (a language of ontologies) and DAML–Logic (a language able to express axioms and rules). It inherits many characteristics of OIL, however, it is less compatible with RDF than OIL.
XSL	Extensible Stylesheet Language	It provides a standard to describe mappings between different terminologies, (a translation mechanism between XML documents).
XOL	Ontology Exchange Language	Simplicity, a generic approach to define ontologies. It has two syntactical variants based on XML and RDF Schema.

Chapter 2

Aspect Based Sentiment Analysis

ABSA framework receive as input a set of texts (product reviews or messages from social media) discussing a specific entity. The systems attempt to identify the main aspects of the entity (storage', camera) and evaluate the average sentiment of the texts for the aspect.

For aspect based sentiment analysis we require two things, sentiment whether it is positive or negative and target for which the opinion is given.

Aspect Extraction:

It extracts the aspects which have been evaluated.Given a set of sentences with particular domain (cellphone), the main function is to distinguish the aspect words occur in the review and return a list having all the aspects. "JBL headphone sound quality is great." Here sound quality is the aspect of the entity JBL headphone. Whenever we have aspect we need to consider which entity aspect belongs to.

Aspect Sentiment Classification:

Check if the opinion is positive or negative for given aspect. In above example, the sound quality is the positive opinion for entity JBL headphone.

Feature selection should be possible by either statistical approach or lexicon based approach. In lexicon-based approach, a seed list of features is manually annotated. Feature selection can treat the words as Bag of Words or n-grams. Feature selection also involves NLP techniques like stop words removal, stemming, etc. The statistical methods can be TF-IDF, PMI etc.

2.1 Approach for Aspect Extraction

Extraction based on frequent nouns and noun phrases:

It finds aspect which is noun and noun phrase from the database for the particular domain. Use part-of-speech(POS) tagger to find noun and noun phrase. In NLP, POS tagging is used. It use some individu tags toal separate adjective, verb, none from the sentences. The disadvantage of this approach is that, it is not necessary that it find the aspect is domain related.

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

[1]

Figure 2.1: POS Tags

Extraction by exploiting opinion and target relations:

It describes the grammatical relationship of sentences using the Stanford type dependencies. Dependencies are represented as directed graph, in which words are the nodes and grammatical relation are edges. Here the verb is selected as the root node. A depen-

dency tree is constructed by the syntax association among a word and its dependents. Using dependency grammar the relation between nodes are define. The parent node is known as the head and its successor are known as modifiers. It uses part-of-speech tag and phrase label. The disadvantage of this approach is that it cannot identify the relation between semantic and syntactic structure. In unigram, occurrence of adjacent word is depends on the occurrence of the previous word this is known as conditional probabilities.

Extraction using supervised learning:

It uses machine learning algorithm like naive Bayes, support vector machine, maximum entropy, conditional random field etc to extract the feature and predict sentiment. One approach to enhance the classification is to develop a feature file that have correct features. For the higher performance in sentiment classifications syntactic and semantic data are used.

Extraction using topic modeling:

Topic modeling is an unsupervised learning method that assumes each document consists of a mixture of topics and each topic is a probability distribution over words. The output of topic modeling is a set of word clusters. Each cluster forms a topic and is a probability distribution over words in the document collection. Topics can cover both aspect words and sentiment words. For sentiment analysis, they need to be separated. Such separations can be achieved by extending the basic model(LDA).

The basic model is LDA which is Latent Dirichlet Allocation, is used. LDA depends on topic distribution differences and word co-occurrences among documents to identify topics and word probability distribution in each topic. The disadvantage of LDA is that the topics are unlabeled, so it will not able to describe a direct relation between topics and a particular aspect of the entity. LDA is designed for the document level. LDA is the popular topic model. LDA is based on the Distributional Hypothesis and the Bag-of-words Hypothesis.

2.2 Types of Features

N-gram Features:

N-grams and their frequency are frequently used. For the frequent word sequences

n-gram is used. For that TF-IDF scheme and word position is considered. There are different types like unigrams, bigrams, and trigrams. Minimum n-gram occurrence is used for feature set. In that 3-grams to 6-grams words are considered. Skip-bigram is used when we have to skip arbitrary gaps between words.

POS-related Features:

Direct usage of POS have not shown any important development in the work. POS labels give positive characteristics to the sentences. Different POS label the features as nouns, verbs, and adjectives.

Lexical Features:

For features, sentiment lexicons or SentiWordNet is used. They use external information to enhance the results of sentiment analysis.

Semantic Features:

Now a days for sentiment analysis distributional semantics are used. They use statistical analysis to give the meaning to the sentences. For classification semantic models are used as source of information.

Syntactic Features:

Use parse trees to generate syntactic information. Syntactic information is used for dependencies and sentence structure for trying to capture features.

Terms and their frequency:

These features are individual words and their n-grams with associated frequency counts. They are also the most common features used in traditional topic-based text classification. In some cases, word positions may also be considered. The TF-IDF weighting scheme from information retrieval may be applied.

Sentiment words and phrases:

Sentiment words are words in a language that are used to express positive or negative sentiments. For example, good, wonderful, and amazing are positive sentiment words and bad, poor, and terrible are negative sentiment words. Most sentiment words are adjectives and adverbs, but nouns (e.g., rubbish, junk, and crap) and verbs (e.g., hate and love) can also be used to express sentiments.

Sentiment shifters:

These are expressions that are used to change the sentiment orientations, from positive to negative or vice versa. Negative words are the most important class of sentiment shifters. For example, the sentence I dont like this phone's camera is negative.

Chapter 3

Related Work

Over the last several years, feature extraction and sentiment analysis have received increasing attention from the research. Ahmad and Najmud Doja[2] proposed an approach called candidate identification and frequent pattern generation. The system uses semantic analysis by using the decision tree classifier and use natural language processing for identifying feature and user FP-growth algorithm to extract feature. The disadvantage is that more memory is required for storing the transaction.Shruti Mishra and Sandeep Kumar[3] use fuzzy pattern tree approach. They compare the performance of original and fuzzified dataset for finding a frequent pattern.Using fuzzified data set it can capable of finding a large number of frequent pattern and have good running time capability. Hui Wang and Jiansheng Chen[4] use two-noun phrase approach for extraction of the feature.Two-noun phrases extract more specific features compared to one-noun phrases. They use three tagging method like CLAWS, NLProcessor, and Lingua::EN::Tagger for checking the accuracy and they find CLAWS tagger give a better result. Gulila Altenbek and Ruina Sun[5] use n-gram method and the experimental result shows that the phrase extraction accuracy is low the alternative for that is the use of basic noun phrase extraction.

A.Jeyapriya and C.S.Kanimozhi Selvi[6] use minimum support threshold to find the frequent aspect and used naive Bayes classifier for sentiment analysis. The disadvantage of the system is that it will not extract relevant feature all the time.Yamamoto, Yamasaki and Aizawa[7] proposed approach service annotation and profiling. They apply IBMWaston relationship to extract POS and used TF-IDF algorithm for computing the relative score. They find the result that the precision without TF-IDF is better.Hamdan,

Bellot and Bchet[8] use the conditional random field to extract feature and use naive Bayes classifier for predict whether the sentence is positive and negative. The disadvantage is that it will not work with unknown words.

Wanying Ding, Zunyan Xiong and Xiaohua Hu[9] proposed the hybrid HDP-LDA model. It uses Dirichlet process to find the aspect using part-of-speech. The advantage of this method is it can automatically determine the number of aspects and it will differentiate actual words from opinion words. Hai Son Le, Thanh Van Le and Tran Vu Pham[10] use GK-LDA for feature extraction and they prove that GK-LDA performs better than the LDA.

Lili Zhao and Chunping Li[11] use an ontology to extract features and use sentiwordnet semi-automatic approach for polarity identification. The disadvantage of this method is they limit the sentiment analysis on the node of the model so its result gives general information about the movie. Mohd Ridzwan Yaakub[12] proposed ontology is too general and used very simple review for the manual extraction of the feature. Meleesa Alfonso and Razia Sardinha[13] use a fuzzy ontology to extract the features and sentiwordnet for polarity. Shein and Nyunt[14] use domain ontology and SVM classifier for sentiment analysis.

Chapter 4

Proposed Method

The main purpose of the proposed system is to give a feature level opinion mining for a particular domain using ontology. Figure 4.1 describes the proposed method.

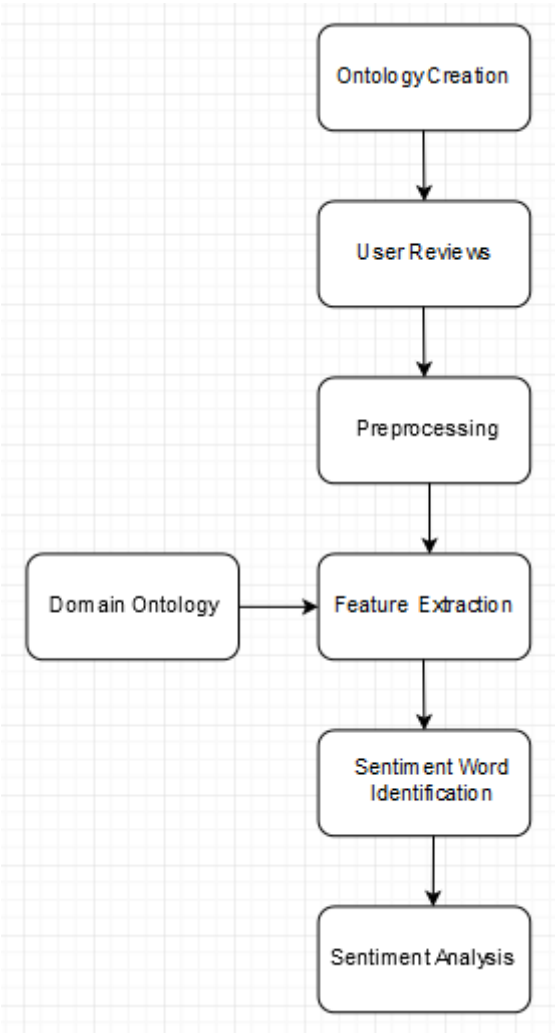

Figure 4.1: Proposed System

4.1 Ontology Creation

An ontology is a formal description of concepts in a domain of discourse (classes), properties of every concept describing various features and attributes of the concept, and restrictions on attributes. The concepts refer to various entities that may be any product or an organization. The use of ontology in feature level opinion mining is to distinguish the domain related features by defining the classes in the domain and giving the relationships between the classes and instances.

Ontology is used to find the domain related features from the domain review. For ontology creation, we can use the existing ontology and extend it as per our requirement or we can build our own ontology. we can use any ontology development tool or java API for ontology creation. Figure 4.2 shows the cell phone ontology.

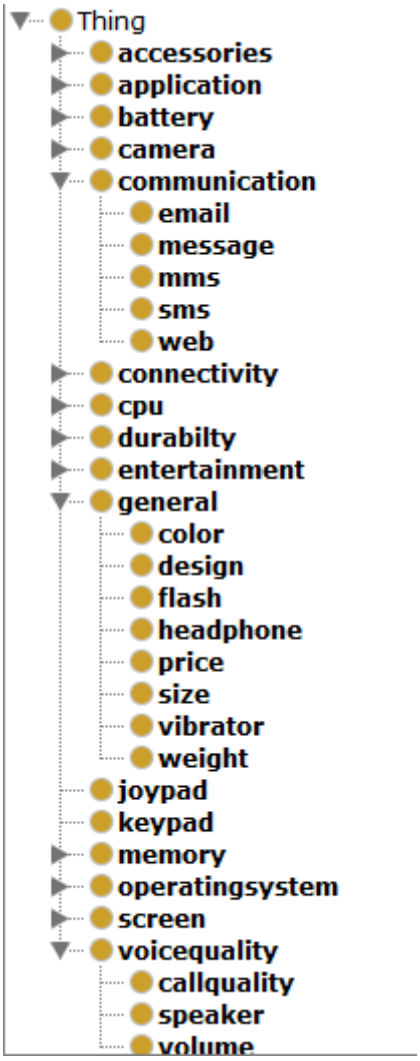

Figure 4.2: CellPhone Ontology

4.2 Preprocessing

Several Natural language preprocessing techniques are available for tokenization, stamming, stop word removal, part-of-speech tagging. In this paper, ontology is used so the only requirement is of POS tagging. POS tagging is used for identification of words as nouns, verbs, adjectives, adverbs, etc. NLTK POS tagger is used to identify the noun.

4.3 Feature Extraction

Domain ontology is used for feature extraction. Identified noun using POS tagger is compared with the concepts of the domain ontology. Here, cellphone ontology is used. Different users write the review and they use different words. For example, a consumer can write cost instead of price, to mention phones rate. So, add synonyms in the domain ontology for better result of feature extraction. In Figure 4.3 users review is available, on that POS tagging is applied and noun are compare with the ontology for feature extraction.

Review	Awesome screen all the greatness of OLED perfect blacks high contrast looks good in direct sunlight it is a pleasure to look at from any distance The resolution is still 800x480 Very light Considering the size of this thing it is very hard to believe When you put the phone into someone else is hand for the first time they usually are confused because they expect it to feel more solid and not so feather weight.
POS tagging	[('Awesome', 'NNP'), ('screen', 'NN'), ('greatness', 'NN'), ('OLED', 'NNP'), ('perfect', 'NN'), ('blacks', 'NNS'), ('high', 'JJ'), ('contrast', 'NN'), ('good', 'JJ'), ('direct', 'JJ'), ('sunlight', 'JJ'), ('pleasure', 'NN'), ('distance', 'NN'), ('resolution', 'NN'), ('800x480', 'CD'), ('light', 'JJ'), ('size', 'NN'), ('hard', 'JJ'), ('put', 'VBD'), ('phone', 'NN'), ('hand', 'NN'), ('time', 'NN'), ('confused', 'VBN'), ('expect', 'VBP'), ('feel', 'NN'), ('solid', 'JJ'), ('feather', 'NN'), ('weight.', 'NN')]
Feature	screen, size

Figure 4.3: Feature Extraction Example.

4.4 Sentiment word Identification

For the identification of positive or negative opinion, sentiment words are used. In this method, adjective words are extracted as a sentiment word. Using POS tagging adjective

can be extracted.

4.5 Sentiment Analysis

Sentiment Analysis means to identify whether the review is positive or negative for a particular feature. Pair the noun and adjective as a feature and its sentiment word. Use sentiment Analyser to analyze whether it is positive or negative.

Chapter 5

EXPERIMENT AND EVALUATION

Apply this proposed method on cellphone review. Take 200 reviews of the cellphone from amazon. Precision and recall are used to measure the performance of the extracted features.

Precision means the words marked as positive are really positive and recall is all the positive words are marked. F-measure is a mean between precision and recall.

$$Precision = \frac{Tp}{Tp + Fp} \tag{5.1}$$

$$Recall = \frac{Tp}{Tp + Fn} \tag{5.2}$$

$$Fmeasure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.3}$$

	Posiitve	Negative
Positive	TP	FN
Negative	FP	TN

Table 5.1

For the 200 Review, a total number of actual extracted feature is 934. Table 5.1 shows

	Tp	Fp	Fn
Features	934	115	391
Sentiments	382	187	291

Table 5.1: Results

it and also it will shows the result for sentiment.

Here, Tp is true positive means the extracted feature and actual feature is same. Fp is false positive means the actual feature is not extracted from the system. Fn is false negative means the feature is extracted but it is not the actual feature. For the feature extraction, Precision is 89.03%, Recall is 70.49% and F-measure value is 78.68%.

For sentiment analysis, Precision is 67.13%, Recall is 56.76% and F-measure value is 61.51%. Here, Tp means the sentiment for the feature is positive and its prediction is also positive. Fp means the sentiment for the feature is negative but the prediction is positive. Fn means the sentiment for the feature is positive but the prediction is negative.

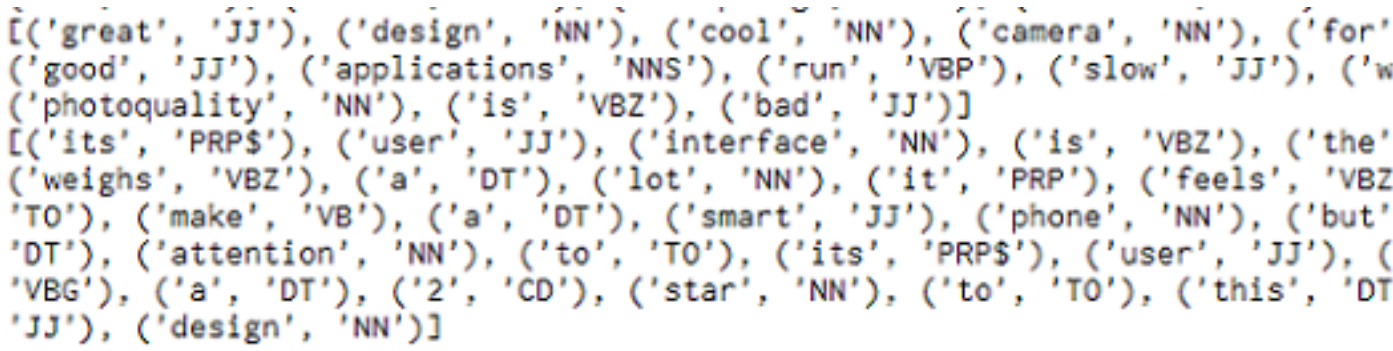

Figure 5.1: POS Tagging

Figure 5.1 describe the result of the POS tagging which is applied to the reviews in the form of the noun(NN), adjective(JJ), verb(VB), etc.

Figure 5.2 shows the features which are extracted from the reviews, using POS tagging noun and ontology.

```
['camera', 'screen', 'phone', 'application', 'phone', 'phone', 'price']
['phone']
['phone', 'camera']
['design', 'camera', 'wifi', 'photoquality']
['phone', 'phone', 'design']
['battery', 'phone', 'phone', 'battery', 'os']
['phone', 'hardware', 'callquality', 'sms', 'phone', 'gps', 'radio', 'phone']
['phone', 'phone', 'camera', 'apps']
['radio', 'camera', 'flash', 'radio', 'phone', 'music']
['gps', 'videos', 'hardware', 'browser', 'battery']
['camera', 'phone']
['camera']
['buildquality', 'phone', 'screen', 'camera']
['camera', 'phone', 'voice', 'speaker', 'gps', 'phone']
['photos', 'camera', 'callquality']
```

Figure 5.2: Features

Figure 5.3 shows the adjectives which are extracted from the reviews, using POS tagging and they are used as sentiment word for sentiment analysis.

```
['actual', 'smooth', 'other', 'excellent', 'nice', 'physical'])
['great', 'good', 'flip', 'long', 'excellent', 'slimmer', 'fine'])
['poor', 'great', 'flash', 'foggy', 'pristine', 'found', 'ovi'])
['real', 'great', 'slow', 'free', 'above'])
['useful', 'light', 'awesome', 'smooth', 'easy', 'excellent', 'loud'])
['flawless', 'good', 'excellent'])
['polished', 'usefull', 'same'])
['perfect', 'great', 'outstanding', 'accurate', 'installed', 'feels'])
['mobile', 'many', 'bad', 'good', 'hang'])
['impossible', 'great', 'call', 'horrible'])
['great', 'slow', 'decent', 'fine', 'lotm', 'disappointed'])
['real', 'good', 'sturdy', 'many', 'unique', 'fine'])
```

Figure 5.3: Adjectives

Figure 5.4 shows the pair of noun and adjective which represent feature and sentiment word respectively.

```
(bluetooth, slow)
(webbrowser, free)
(apps, above)
(phone, useful)
(connectivity, light)
(email, awesome)
(camera, smooth)
(callquality, flawless)
(wifi, good)
(phone, excellent)
(camera, polished)
(usb, usefull)
(phone, perfect)
(size, great)
(touchscreen, outstanding)
(wifi, accurate)
```

Figure 5.4: Pair Of Feature and Sentiment Word

Figure 5.5 describe the final output. In that, the first line is of extracted feature. The second line shows the pair of extracted feature and sentiment word. The fourth line shows the sentiment whether it is positive or negative.

```
['camera', 'phone']
(camera, perfect)
(phone, loud)
{'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.5719}
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
['photos', 'camera', 'callquality']
(photos, past)
(camera, outstanding)
(callquality, excellent)
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
{'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.6124}
{'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.5719}
```

Figure 5.5: Final Result

Chapter 6

Conclusion

Ontology is an idea which describes system in the form of knowledge and semantics. It expects to get information and provides an understanding of information in the domain. Opinion mining is a task, which analyzes opinions given by users for different features, and determines whether these opinions are positive, neutral or negative.

In this system, Ontology is used for the feature level sentiment analysis for the cell phone review. It is used for both producer and consumer. The producer can improve their services and product feature and it will help the consumer to make the decision about what to buy based on the features they like. First cellphone ontology is created for feature extraction. Feature and sentiment word pair is used for sentiment analysis.

For the future work develop the system which will automatically update the ontology and take two different domain ontology and try to extract features related to the particular domain.

Bibliography

[1] B. Liu, *Sentiment analysis and opinion mining*, Synthesis lectures on human language technologies, vol.5, no.1, pp.1-167, 2012.

[2] T.Ahmad and M.N.Doja, *Opinion mining using frequent pattern growth method from unstructured text*, pp. 92-95, 2013.

[3] S.Mishra, D.Mishra, and S.K.Satapathy, *Fuzzy pattern tree approach for mining frequent patterns from gene expression data*, vol.2, pp.359-363, 2011.

[4] H.Wang and J.Chen, *Extracting two-noun phrases from customer reviews*, pp.1-4, 2009.

[5] G.Altenbek and R.Sun, *Kazakh noun phrase extraction based on n-gram and rules*, pp.305-308, 2010.

[6] A.Jeyapriya and C.K.Selvi, *Extracting aspects and mining opinions in product reviews using supervised learning algorithm*, pp. 548-552, 2015.

[7] M.Yamamoto, T.Yamasaki, and K.Aizawa, *Service annotation and profling by review analysis*, pp.357-364, 2016.

[8] H.Hamdan, P.Bellot, and F.Bechet, *Supervised methods for aspect-based sentiment analysis*, 2014.

[9] W.Ding, X.Song, L.Guo, Z.Xiong, and X.Hu, *A novel hybrid hdp-lda model for sentiment analysis*, pp.329-336, 2013.

[10] H.S.Le, T.Van Le, and T.V.Pham, *Aspect analysis for opinion mining of Vietnamese text*, 2015.

[11] Zhao, Lili, and Chunping Li, *Ontology based opinion mining for movie reviews*, International Conference on Knowledge Science, Engineering and Managemen, 2009.

[12] Yaakub, R.M, Li and Feng Y, *Integration of Opinion into Customer Analysis Model*, IEEE International Conference on e-Business Engineering, pp.90-95, 2011.

[13] Alfonso,Ms Meleesa and Ms Razia Sardinha, *Ontology based Aspect level Opinion Mining*, International Journal of Engineering Sciences & Research Technology.

[14] Shein, Khin Phyu Phyu and Thi Thi Soe Nyunt, *Sentiment classification based on Ontology and SVM Classifier*, IEEE Communication Software and Networks, 2010.

[15] Freitas, Larissa A, and Renata Vieira, *Ontology based feature level opinion mining for portuguese reviews*, 22nd International Conference on World Wide Web. ACM, 2013.

[16] Hazman, Maryam, Samhaa R.El-Beltagy and Ahmed Rafea, *A survey of ontology learning approaches*, 2011.

[17] Penalver-Martinez, Isidro, et al, *Feature-based opinion mining through ontologies*, Expert Systems with Application, 2014.

[18] Ali, Farman, Kyung-Sup Kwak and Yong-Gi Kim, *Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification*, Applied Soft Computing, 235-250, 2016.

[19] Kontopoulos, Efstratios, et al, *Ontology-based sentiment analysis of twitter posts*, Expert systems with applications, 4065-4074, 2013.