I.J. Intelligent Systems and Applications, 2016, 7, 55-64
Published Online July 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2016.07.06

An Automated Real-Time System for Opinion Mining using a Hybrid Approach

Indrajit Mukherjee
Department of Computer Science and Engineering, BIT Mesra, Ranchi, INDIA
E-mai: imukherjee@bitmesra.ac.in

Jasni M Zain
Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, Malaysia
E-mai: jasni@ump.edu.my

P. K. Mahanti
Department of CSAS, University of Brunswick, New Brunswick, Canada
E-mai: pkmahanti@yahoo.co.in

Abstract—In this paper, a novel idea is being presented to perform Opinion Mining in a very simple and efficient manner with the help of the One-Level-Tree (OLT) based approach. To recognize opinions specific for features in customer reviews having a variety of features commingled with diverse emotions. Unlike some previous ventures entirely using one-time structured or filtered data but this is solely based on unstructured data obtained in real-time from Twitter. The hybrid approach utilizes the associations defined in Dependency Parsing Grammar and fully employs Double Propagation to extract new features and related new opinions within the review. The Dictionary based approach is used to expand the Opinion Lexicon. Within the dependency parsing relations a new relation is being proposed to more effectively catch the associations between opinions and features. The three new methods are being proposed, termed as Double Positive Double Negative (DPDN), Catch-Phrase Method (CPM) & Negation Check (NC), for performing criteria specific evaluations. The OLT approach conveniently displays the relationship between the features and their opinions in an elementary fashion in the form of a graph. The proposed system achieves splendid accuracy across all domains and also performs better than the state-of-the-art systems.

Index Terms—Opinion Mining, Sentiment Analysis, Feature Extraction, Twitter Data Analysis, Graph based Sentiment Analysis, Data Extraction.

I. INTRODUCTION

The recent boom of social networking sites and blogs provide a lot of customer based information [15]. This sets the tone for a dynamic feedback system which is of importance not only to the companies developing the products, but also to their rivals and several others potential customers [25]. Opinion Mining comes to the aid for utilizing this feedback system [26].

Opinion Mining out performs in the task of intercepting and interpreting the customer opinions. It has the sole task of extracting opinion words for a certain product or features and assigns them as positive or negative, suggesting the impression of each product has left on its customers and therefore whether they are recommending or not. The opinions regarding any particular product in a review tend to be a mixed opinion about various features, some positive and some negative. Thus the feature specific opinion matters more than the overall opinion. Furthermore any comprehensive customer who has used a certain collection of products gives a comparative analysis of products with his certified grading opinions. Such reviews can in turn provide a hierarchical relationship among the products.

In this paper, we propose a method which will simply represent the features and opinions in the form of a One-Level-Tree (OLT) following the concept of *"Feature taking up the Root position and all the opinions are depicted as Child Nodes"*. Dependency Parsing is used to capture any associations between the features and their intended opinions. Double Propagation is implemented to extract more features and correspondingly more related opinions. The dictionary based approach is used to expand the opinion lexicon. The three approaches have been identified as Double Positive Double Negative (DPDN), Catch-Phrase Method (CPM) and the Negation Check (NC) methods. We apply our approach to the data-set obtained from twitter and also check against other existing systems and the state-of-the-art system to find out some very good exciting results.

The organization of the paper is as follows:

Section 1 presents the interest and motivation towards this work. Section 2 shows the related works and the health of the existing system. Section 3 describes the proposed work. Section 4 gives the procedure for Data Extraction from Twitter. Section 5 gives the Data Filtering. Section 6 represents the Feature Specific Opinion Mining. Section 7 includes the results of the system created and its utility. Section 8 displays the Experimental Evaluation of the proposed approach to the other existing systems. Section 9 gives the conclusions and effective ideas for future scope of work followed by references.

II. Related Work

Hatzivassiloglou and McKeown [1] gave the initial idea for determining adjective polarities or orientations (positive, negative, and neutral). The method involves the use of small seeded lexicons for the initial assessment and then predicting orientations of adjectives and other opinions by detecting pairs of such words conjoined by conjunctions such as *and/or/but/where as* in a large document set. The weakness of this initial approach is that it relies heavily on the list of conjunctions used and the set of conjunction relations. Wiebe [2] and Wiebe et al. [3],both the research papers proposed an approach to search obscured adjectives using the findings of word clustering according to their distributional similarity. Turney and Littman [4] proposed an approach to compute the point wise mutual information (PMI) of the target feature with each seed positive and negative term as a measure of their semantic relationship. In the dictionary-based approach, the initial idea of taking advantage of WordNet to construct a network of opinions by connecting pairs of synonymous words found in the data-set [5]. The semantic polarity or orientation of each word is decided by the shortest paths from itself to two seed words good and bad which are chosen as illustrators of positive and negative orientations. Takamura, Inui & Okumura [6] also exploits the idea of fetching obscure information with the help of dictionaries. This method constructs a lexical network by linking two words if one appears in the gloss of the other. The works of Hu and Liu [7] and Kim and Hovy [8] are simpler as they efficiently used synonyms and antonyms to simply expand the opinion lexicon.

Most of the previous work relates to Sentiment Analysis and classification of tweets into positive and negative buckets. A good deal of work in sentiment analysis has been done in [23], [28] and [24]. Go, et. al. [28] were one of the earliest to extend machine learning approaches to sentiment analysis of Twitter data. They use tweets with emoticons to train their model. Tweets containing positive emoticons correspond to positive sentiment and tweets containing negative emoticons correspond to negative sentiment. The authors test their data using manually labeled tweets and report highest accuracy using SVM trained with unigram features. Pak and Paroubek [24] work on classifying tweets into subjective vs. objective. Subjective tweets are those that carried a sentiment and objective tweets are those that did not. They use the same methodology for training as [28]. They introduce a method of increasing accuracy by disregarding common n-grams. The authors report better performance with bigrams over unigrams and trigrams. The authors explain that this might be because bigrams provide good balance between coverage and ability to capture sentiment expression patterns. The authors in [23] use manually annotated data to train the classifier. They claim that tweets collected using queries tend to be biased. They also introduce twitter sentiment features along with unigrams to train their classifier and report improvement in accuracy.

III. Proposed Work

The proposed work is semi-supervised methods in the presence of domain knowledge since unsupervised methods do not perform well. However, semi-supervised

Methods are needs some labeled examples. The work was exercised on three products mainly {Mobiles, Cars, & Bikes}. Each product was further sub-divided into categories for example; Mobiles were classified into 4 categories, like “Price less than 5K”, “Price 5K-10K”, “Price 10K-15K”, and “Price 15K-20K”.For each individual category a total of 50 items were considered. The work can be broadly classified into three distinct modules:-

1. *Data Extraction from Twitter –*

We obtain real-time data in the presence of Domain Knowledge.

2. *Data Filtering–*

Data Filtering is done to convert obtained data to usable data.

3. *Feature Specific Opinion Mining –*

Relation Extraction and in-depth feature extraction is done with the help of Dependency Parsing, Double Propagation & a number of newly proposed methods.

IV. Data Extraction from Twitter

Twitter has introduced a new standard called Open Authentication (OAuth) where only authenticated requests are allowed to extract data. As mentioned above, the real-time data-set is obtained on three products namely {Mobiles, Cars, & Bikes}.Each product was divided into categories. For each individual category a total of 50 items were considered. Hence a total of 3 products each have 4 categories and each category having 50 different items. Total of 4,00,000 tweets were evaluated.

The data obtained from twitter is extracted from its source code. As shown in figure 1, the extracted data contains information mostly about attributes which will not be required in our work. Hence such data obtained will have to be filtered to meet our requirement.

{'contributors': None, 'truncated': False, 'text': "@RikiRachtmanThat's why I switched the the Samsung Galaxy S4. Better phone, no drama. Good luck!!!!!", in_reply_to_status_id ': 407308499039686656L, 'id': 40730917814364979 2L,'favorite_count': 0, 'source': 'web', 'retweeted': False, 'coordinates': None, 'entities': {'symbols': [], 'user_mentions': [{'id': 67978376, 'indices': [0, 13], 'id_str': '67978376', 'screen_name': 'RikiRachtman', 'name': 'RikiRachtman'}], 'hashtags': [], 'urls': []}, 'in_reply_to_screen_name': 'RikiRachtman', 'in_reply _to_user_id': 67978376, 'retweet_count': 0, 'id_str': '40730917 8143649792', 'favorited': False, 'user': {'follow_request_sent': False, 'profile_use_ background _image': True, 'default_profile_image': False, 'id': 349130261, 'verified': False, 'profile_text_color': '362720', 'profile_image_url _https':'https://pbs.twimg.

Fig.1. Result of Data Extraction

V. Data Filtering

The Data obtained from Twitter has many impuissance`s and disadvantages. Twitter allows extracting data but the extracted data is in its Source Code. The Source Code contains many redundant piece of extraneous data or information like *contributors, truncated, text, in-reply-to-status-id, id, favorite-count, source, retweeted, etc*. The only thing required here is to remove any unnecessary Tags and Hash-Tags. A Hash-Tag is a word or a group of words or without spaces in between collection of words making a single phrase prefixed with the symbol "#" making a significant pattern of metadata. The Tags are used for tagging other people into the conversation.

Both of which are not required in the analyzing of opinions. Since the initiation of the Social Networking Sites, the art of conversation shifted from speaking to typing. But typing consumed a huge amount of time which wants to waste. Hence the idea of abbreviations came into being during conversations. It was seen that some of the texts were very common during a conversation such as *GOODNIGHT*. So *GOODNIGHT* became *GN*. For a proper Opinion Mining, an Acronym Lexicon had to be created. The Lexicon as of now consists of 1500 abbreviations. But such a list can be complete with the ever increasing trend of new acronyms generation every day.

Figure 2, 3 & 4 are results of Data Filtering for different purposes where each step is done after the other. The extraction data from twitter had many attributes but we required only the text part or the so called “tweet-text” part. Hence by scanning the data only for “tweet-text” gave figure 2. The results of obtaining only the texts can be seen in figure 2. It can be noticed that the data extracted still had many impurities such as the sender of the tweet as well as the number of people in the conversation added by tags and also grouped conversation pivots represented by hash tags. So they had to be eliminated because they served no purpose, resulting in figure 3. Recent trends show that most people save time in conversation by using acronyms. This is a new approach taken in this paper to convert acronyms to full grown words or phrases to extract sentiments embedded in those acronyms.

- ❖ @Riki Rachtman That's why I switched the Samsung Galaxy S4. Better phone, no drama. Good luck!!!!!
- ❖ @Rachael Evans Crazy, Confused, Fun, Mad, Happy, Silly, Goofy, Smart and more is Samsung S4. lolim going crazy.
- ❖ u'RT @bathroomsupa: The #Competition heating up in our #freetoenter #Giveaway for a #Samsung Galaxy S4

Fig.2. Extracting only Tweets from the data obtained from Twitter

- ❖ @Riki Rachtman That's why I switched the Samsung Galaxy S4. Better phone, no drama. Good luck!!!!!
- ❖ @Rachael Evans Crazy, Confused, Fun, Mad, Happy, Silly, Goofy, Smart and more is Samsung S4. lolim going crazy.
- ❖ u'RT @bathroomsupa: The #Competition heating up in our #freetoenter #Giveaway for a #Samsung Galaxy S4

Fig.3. Removal of TAGS and HASH TAGS

- ❖ That's why I switched the Samsung Galaxy S4. Better phone, no drama. Good luck!!!!!
- ❖ Crazy, Confused, Fun, Mad, Happy, Silly, Goofy, Smart and more is Samsung S4. laugh out loud I am going crazy.
- ❖ You are right The Competition heating up in our free to enter Giveaway for a Samsung Galaxy S4

Fig.4. Removal of ACRONYMS

VI. Feature Specific Opinion Mining

One of the main issues in Opinion Mining is to realize the relations between the words in the reviews. This is done through Dependency Parsing. Double Propagation has been utilized here only for expanding respective Domain Sentiment Lexicons within the review. The Dictionary-based approach is based on finding antonyms and synonyms of opinions extracted from reviews. The DPDN Method has been proposed to encounter the comparative structure of multiple features in a single review. The CPM Method has been proposed to capture the overuse of phrases & catchphrases. The NC Method has been proposed as an extended version for ascertaining the negative opinions in complex reviews.

6.1 Dependency Parsing

The relationship among the words is realized through Dependency Parsing. It has been identified that there exists two kinds of associations between the words in a sentence that links them together to form a logically ordered review i.e. Direct Neighbor Relation and Indirect Relation. We wish to propose another kind of relation i.e. Transitive Relation. This is a unique case of opinions clinging to a false positive feature. The hypothesis for Transitive Relation is as follows:

"If in any review there is a supposed feature, which is not the true feature, to which most of the opinions are related then the distance between the supposed feature and other features present is calculated. That feature, with which its distance is minimum, is selected as the true feature".

Considering the review *"Nokia Asha 502 is the best available budget phone in the country"*. By implementing only the Dependency Parsing Relations as shown in Figure 5, it can be seen that *"Phone"*&*"Nokia Asha 502"* are the features and the major opinions are {Best, Available, Budget & Country}. Both *"Phone"*&*"Nokia Asha 502"*are now eligible to form OLT`s. All of the relations are formed between the feature "phone" and the opinions. But *"phone"* is not the true featured item in this review. *"Nokia Asha 502"* is the intended true item of this review.

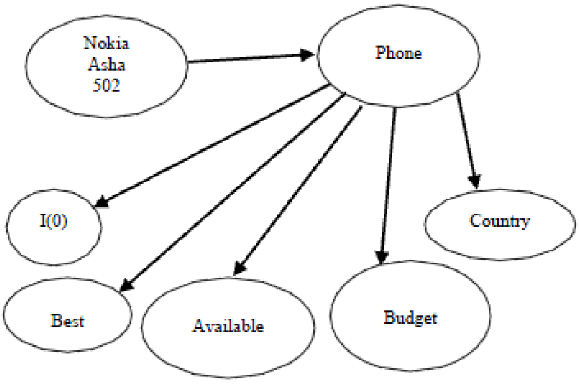

Fig.5. Depiction of Transitive Relation

6.2 Double Propagation

Double Propagation has been implemented only for expanding respective Domain Sentiment Lexicon. The Double Propagation method first utilizes the sentiments or opinion words and features already extracted initially using the Opinion Lexicon. After that it then employs the extracted sentiment or opinion words and features to search for new sentiment or opinion words and new features. The newly extracted sentiment or opinion words and new features found are then employed to extract and search for more sentiment or opinion words and new features in the same manner. The process keeps on continuously searching and extracting new opinion words and new features until no more sentiment or opinion words can be found and added.

Utilizing Double Propagation four types of extraction of feature & opinions can be done separately considering each review individually :-

- ❖ *Extraction of sentiment words using sentiment words.*
- ❖ *Extraction of features using sentiment words.*
- ❖ *Extraction of sentiment words using features.*
- ❖ *Extraction of features using features.*

6.3 The Dictionary-based Approach

The Dictionary-based approach is used on finding antonyms and synonyms of opinions from the reviews. This approach employs WordNet. WordNet consists of a set of seed opinion words. So when any particular opinion has to be checked then it is evaluated with the opinion lexicon of WordNet.

/* Algorithm for the Dictionary() Procedure */

```
For each word W in review R do
    Compare W with OL and find its polarity
end for
   For each word W in review do
   Search W in WordNet Lexicon
   Assign the synonyms found to the network W
   with the same polarity.
   Assign the antonyms found to the network (-W)
   with the opposite polarity.
end for
end procedure
```

6.4 The DPDN Method

DPDN stands for "Double Positive & Double Negative". It has a simple functionality of comparing two or more item features in any customer review. Also such reviews have been found to be totally positive or totally negative. Consider the reviews *"The Samsung Galaxy Series is good but the Nokia N Series is better"* and the review *"I prefer Audi to BMW"*. Both the reviews talked about the personal preference of the customer but also gave a comparison between two different product items or features where one product is preferred to the other. The DPDN method can be extended to properly define the *hierarchical relation* among the features according to the review & also use *multi-polarity sentiment values*.

/* Algorithm for the DPDN() Procedure */

```
 Initialize the feature set FS[i] where i := 1….F
 Check the connectivity among the F features in FS
 If the connector belongs to Positive ConjunctionList
 then
     Assign the same polarity to each of the opinions
 end if
 If the connector belongs to Negative Conjunction_List
   then
      Assign Multi-Polarity values in ascending order.
 end if
end procedure
```

6.5 The CPM Method

Recent trends have shown that people generally converse with little words and more abbreviations.CPM stands for *"Catch-Phrase Method"*. Two new Lexicons were created each for the Catch-Phrases called the Catch-Phrase Lexicon (CPL) and the Extended Phrase Lexicon (PL).The CPM procedure can also be extended to be used in reviews without any specific opinions such that will be considered positive. E.g. consider the review *"This is the mother of all bikes"*.

/* Algorithm for the CPM() Procedure */

```
Check each word with its neighbor like W and W+1
   If any pair matches totally or even partially with any
element in the CPL
then
         Assign the corresponding polarity value to the
words(W,W+1,……)  from CPL
end if
Check each word with its neighbor like W and
W+1   Again
If any pair matches totally or even partially with
any element in the PL
then
       Assign the corresponding polarity value to the
words(W,W+1,……)  from PL
end if
end procedure
```

6.6 The NC Method

The NC method is an extended form of *Negation Checking*. The NC method will be fully implemented through the use of the DIT relations. The DIT relations stand for Direct Relations, Indirect Relations and Transitive Relations. Consider the review *"Not that Motorola Moto G is a bad product but it can be better"*. The prevailing methods will compare the distance and evaluate that Motorola Moto G is being opinionated as *"a bad product"*. Yet the truth is entirely different. So with the help of the NC method such problems can be rectified.

/* Algorithm for the NC() Procedure */

```
 Check for Negations in Direct relations i.e. among
   neighbors
  If found then
   Assign a negative polarity
 end if
  Check for Negations in Indirect relations i.e. among
   non-neighbors
    If found then
      Couple the group of opinion words
      expressing  the negation.
      Assign a negative polarity
   end if
end procedure
```

6.7 The Proposed OLT Approach

The very basic aim of this approach is to provide "More Simplicity & More Accuracy". OLT is a coined term representing a One-Level-Tree. OLT is the most simplistic approach through which features and opinions both can be represented together in an elementary relationship. The hypothesis for the OLT Structure is as follows:

"Feature taking up the Root position and all the opinions are depicted as Child Nodes".

/* Algorithm for OLT */

```
Initialize  the F OLT`s  for i := 1….F
      Call  Dependency_Parsing() procedure
For each feature F in FS  where j := 1….F  do
Make F the OLT  head for its respective FS[j]
 For each opinion word  W do
Assign W to its respective FS[j]
        Call  Double_Propagation() procedure
 For each new feature extracted  NFS[M]
 For k :=   1….M  do
Create new OLT  with FS[k]
 For each new extracted opinion word  NW do
          Assign NW to its respective FS[k]
Call  the Dictionary() procedure to expand the OL of the
system.
 Call  DPDN()  procedure,
 Call  CPM() procedure and the NC()  procedure to modify
the present OLT`s  FS[j+k]wherever  necessary.
Comparing  each OLT`s  FS[j+k]  with the OL assign the
polarity values.
 Compute  the overall  Sentiment of the Review  R.
```

Consider the review "I want to ride the Audi A8 but I am not so sure about the Volvo V60"

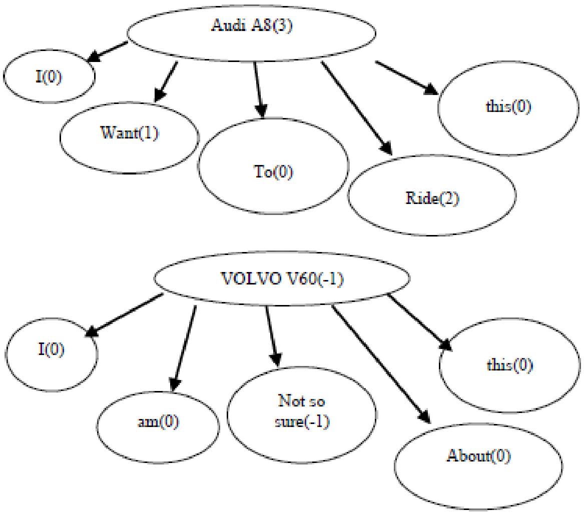

Fig.6. OLT`s for Audi A8 & Volvo V60

Figure 6 represents the final OLT structure and the final sentiment polarity of both the product features. The values in parenthesis represents the sentiment values of each of the opinions and summation of all the sentiment values gives the overall sentiment verdict of the product feature. As can be seen the structure is very simple and very simple to understand. Firstly with the help of Dependency Parsing, the associations between the words are found. Relations are extracted among the opinions and between the features and the opinions. This simply divides the opinions into two groups such that each group of opinions corresponds to a single product feature. The OLT is constructed for each of the feature. Now the opinions are checked with the Opinion Lexicon, the Catch-Phrase Lexicon and the Extended Phrase Lexicon to find the sentiment polarity of each of the opinions found here and also their corresponding polarity value. The opinions with 0 polarity value represents neutral words and hence a neutral polarity value. The Lexicon considered the preposition, vowels, is, was, are, have, has etc verbs acts as neutral words Polarity. Finally the polarity values of all the opinions are summarized to evaluate the overall polarity value of the product feature.

VII. Results for the Designed System

Table 1a and Table 1b represent the database of mobiles formed for category "Price Less than 5K" with having feature sets Primary Camera, Internal Memory, RAM, Card Slot, Battery, Size, Display, OS. Due to large feature sets of the mobile, table is shown in two part Table 1a, Table 1b. As depicted above each product was divided into 4 categories and each category had 50 items in it. For all the 50 items in each category for each product 500 tweets were extracted. The data was obtained by evaluating the opinions on each corresponding basic attribute as shown in the figure. But those 500 tweets were proportionally lessened to show results up to a standard of 100 tweets for each category product. This figure is a one of the many database obtained in our work.

Figure 7 & 8 have been used to depict the query system which can be used in multiple ways from the data obtained. Such data all the mobiles can be compared against a single feature as shown in figure 8. Also all the features of any product can be shown in a comparative way based on customer preferences.

The procedure applied here was tested by performing the experiments in stages such that to find to find best possible solution route. Five methods had to be implemented namely, Dependency Parsing, Double Propagation, DPDN, CPM and the NC method.

Table 1a. Database of Mobiles for Category Price Less than 5K

Mobile Phones	Primary Camera		Internal Memory		RAM		Card Slot		Battery	
	P	N	P	N	P	N	P	N	P	N
Nokia asha 502	28	8	14	16	3	25	34	16	28	10
Micromax Bolt A67	30	13	41	8	11	9	14	15	12	47
LGOptimus l3 e400	35	14	11	5	25	12	22	10	18	5
Samsung Galaxy Star	30	13	28	20	25	10	18	14	20	39
Micromax Canvas A63	18	12	44	8	46	3	25	9	47	13
Nokia 208	16	15	28	16	12	10	33	11	28	2
Blackberry Curve8250	16	14	19	23	28	17	37	8	35	15
NokiaAsha 305	36	14	6	36	15	23	29	7	34	3
Nokia X2-02	33	11	10	29	18	27	31	9	45	10
Nokia X2-01	31	9	13	26	19	21	34	12	40	11
Nokia Asha 210	25	11	13	18	10	33	28	31	32	21
Nokia Asha 230	28	9	15	19	15	27	26	34	38	20
Nokia x2	34	12	10	23	21	26	35	13	58	14
Nokia 301	33	17	11	21	14	29	13	19	16	24
Micromax Belt A35	38	12	23	25	24	27	12	17	21	22
Micromax Belt A 27	21	26	27	28	23	26	15	14	20	29
Samsung DousC3321	34	18	33	12	27	20	13	15	14	12
Huawei Ascend T2100	25	21	12	14	23	12	11	14	15	10
Micromax Bolt A 62	23	14	29	26	21	22	13	16	21	27
Samsung c3322	33	10	21	25	22	15	13	6	17	14
Samsung star 55230	38	12	29	10	27	13	26	9	11	25
Samsung Neo c3262	9	35	19	23	15	18	26	21	10	29
Micromax Belt A36	26	15	27	21	20	24	12	16	24	26
Nokia Asha 308	30	13	16	23	15	24	16	19	26	14

Table 1b. Database of Mobiles for Category Price Less than 5K

Size		Display		OS		Rating in class	Overall User Liking (%)
P	N	P	N	P	N		
17	18	16	19	21	11	4	54
21	25	15	12	24	9	3	32
26	15	28	3	32	10	3.5	21
35	25	21	5	38	16	5	52
22	16	38	6	24	12	5	42
18	23	12	26	34	24	4	28
21	7	32	18	28	6	5	55
13	9	38	18	28	43	4	35
28	13	20	25	30	15	4	37
22	6	19	23	33	18	4	39
23	12	21	25	19	12	4	43
28	14	26	29	13	9	3	33
26	5	22	27	37	19	4	38
17	25	13	26	14	7	2	18
14	10	26	28	27	14	3	26
12	12	23	24	23	15	3.5	30
23	11	18	7	17	13	3.5	32
9	7	17	12	15	11	3.5	25
13	15	21	26	28	13	3	29
19	8	18	12	16	9	2	21
16	19	13	20	22	6	4	37
22	26	31	14	11	9	3	22
15	17	35	12	24	14	3.5	33
20	12	28	19	23	13	4	36

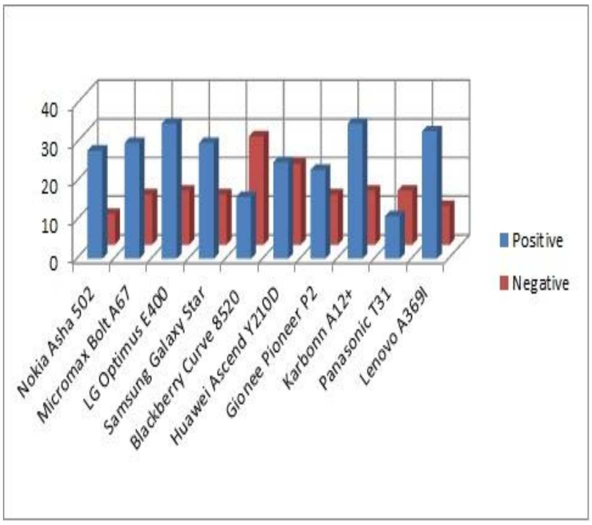

Fig.7. Screenshot for Mobile Comparison for a Feature, like “Camera”

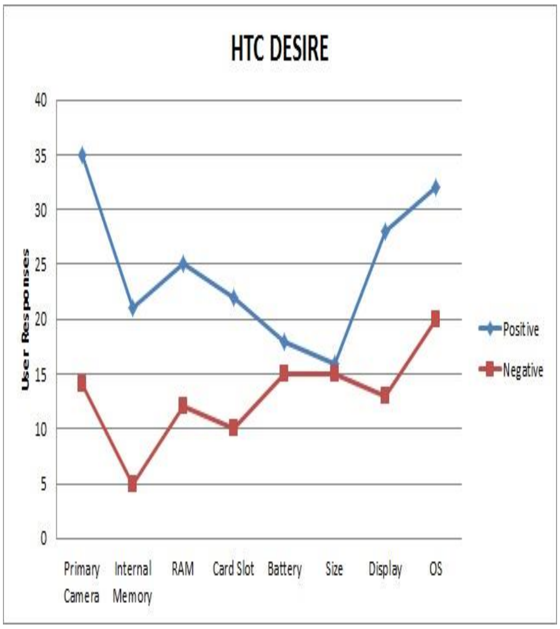

Fig.8. Displaying Features of a Mobile like HTC Desire

Table 2. First Approach

Methods Applied	Mobiles	Cars	Bikes
Dependency Parsing	53.24 %	52.76 %	54.88 %
Double Propagation	68.39 %	65.54 %	69.73 %
DPDN	71.15 %	70.01 %	73.84 %
CPM	74.37 %	73.22 %	77.31 %
NC	78.92 %	77.67 %	80.33 %
System Accuracy	**78.92 %**	**77.67 %**	**80.33 %**

System Accuracy = 78.97 %

Table 3. Second Approach

Methods Applied	Mobiles	Cars	Bikes
Dependency Parsing	53.24 %	52.76 %	54.88 %
Double Propagation	68.39 %	65.54 %	69.73 %
CPM	70.15 %	68.23 %	71.56 %
DPDN	72.42 %	71.22 %	73.23 %
NC	75.68 %	74.59 %	75.73 %

System Accuracy = 76.33 %

Table 4. Third Approach

Methods Applied	Mobiles	Cars	Bikes
Dependency Parsing	53.24 %	52.76 %	54.88 %
Double Propagation	68.39 %	65.54 %	69.73 %
NC	75.15 %	77.84 %	78.71 %
DPDN	80.17 %	81.62 %	80.41 %
CPM	83.75 %	84.25 %	83.78 %

System Accuracy = 83.92 %

Table 2, 3 & 4 represents three different approaches taken here to evaluate that which method applied what gives the best results. It can be seen that approach 3 shown in Table 4 gives the best results. This might be different for a different data-set but here this approach works best.

Table 5. Comparative Analysis among Various Other Approaches

System	Accuracy
CFACTS-R	80.54 %
CFACTS	81.28 %
FACTS-R	72.25 %
FACTS	75.72 %
JST	76.18 %
Rule based system using a graph	80.98 %
Proposed System using OLT	83.92

VIII. Experimental Evaluation

The average accuracy of the 10-fold cross-validation result of each configuration is to be considered in this experiment. The training data were separated into ten folds, and the system used 90% of the data as the training set and the other 10% as the test set. Accuracy is defined as the proportion of true positive, true negatives and true neutrals (true results) from all the given data.

Table 5 gives the comparisons were also made with a state-of-the-art system which is, CFACTS formulated by Lakkarajuet. *Al* [9]. The CFACTS system claims to have 100% topic purity in feature extraction which means its feature extraction accuracy cannot degrade its sentiment evaluation accuracy. To carry out this comparison another data-set consisting of 500 reviews extracted from the dataset used by Lakkaraju*et. Al*[9]was used. This data-set contained data from 3 domains *laptops, camera and printers.*

Table 5 gives a comparison among the existing systems with our proposed technique. It is clear from the comparison that the proposed algorithm shows better accuracy. This analysis was based on using the data-set as mentioned above on 3 domains. The state-of-the-art system CFACTS is also based on domain dependency knowledge as similar as our system.

IX. Conclusion and Future Scope of Work

Through this paper a sincere effort has been given to successfully extract complicated and mixed opinions corresponding to the known feature. The main gist and its corresponding improvement is as follows:

1. In this paper, many new methods have been proposed and also a new relation, i.e. Transitive Relation, has been proposed. The DPDN method can be employed to exploit the hierarchical relations among the product items. The CPM method is more than useful according to the current trends in social networking blogs or sites. The NC method excels in the aggregation of negative and neutral words which in turn becomes a negative opinion and further it adds to the OLT.
2. Most important thing to be noted about the OLT is its simplicity - in architecture and in meaning. Once created it can be modified, merged and even collapse without any difficulty. The OLT can further be utilized in multi-polarity usages
3. The proposed system has performed better than most of the contemporary domain dependent systems including the state-of-the-art CFACTS system.
4. Extracting tweets real-time isn't really the problem. But filtering that data into a standard for analyzing on it is tough. The main issue is with the kind of language now-a-days people use to convey their messages. Hence it's a big issue.
5. The Lexicons used here is still not suited to understand sarcasm and subtlety. Sarcasm as we all know is saying something but whose hidden meaning is entirely different or just the opposite. Subtlety is way of expressing one`s expressions. This nevertheless remains a tough task.

References

[1] Hatzivassiloglou, Vasileios and Kathleen R. McKeown, "Predicting the semantic orientation of adjectives", In Proceedings of ACL'97, pages 174–181. Stroudsburg, PA, 1997.

[2] Wiebe, Janyce, "Learning subjective adjective from corpora.", In Proceedings of AAAI'00, pages 735–740, 2000.

[3] Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin, "Learning subjective language", Computational Linguistics, 30(3):277–308, 2004.

[4] Turney, Peter D. and Michael L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association", ACM Transactions on Information System, 21(4):315–346, 2003.

[5] Kamps, Jaap, Maarten Marx, Robert J. Mokken, and Maarten de Rijke, "Using Wordnet to measure semantic orientation of adjectives", In Proceedings of LREC'04, pages 1115–1118, 2004.

[6] Takamura, Hiroya, Takashi Inui, and Manabu Okumura, "Extracting semantic orientations of words using spin model", In Proceedings of ACL'05,pages 133–140, 2005.

[7] Hu, Mingqing and Bing Liu, "Mining and summarizing customer reviews", In Proceedings of SIGKDD'04, pages 168–177, 2004.

[8] Kim, Soo-Min and Eduard Hovy, "Determining the sentiment of opinions", In Proceedings of COLING'04,pages 1367–1373, 2004.

[9] Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya and Srujanaerugu,. "Exploiting Coherence for the simultaneous discovery of latent facets and associated sentiments", SIAM International Conference on Data Mining (SDM),2011.

[10] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews", KDD '04:Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[11] Chen Mosha, "Combining Dependency Parsing with Shallow Semantic Analysis for Chinese Opinion-Element Relation Identification", IEEE, pp.299-305, 2010.

[12] Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu, "Phrase Dependency Parsing for Opinion Mining", EMNLP '09 Proceedings of the 2009 Conference on Empirical Methodsin Natural Language Processing, Volume 3, 2009.

[13] Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang, "Mining Product Reviews Based on Shallow Dependency Parsing", SIGIR09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.

[14] M. Mathioudakis and N. Koudas, "Twitter monitor: Trend Detection over the Twitter Stream", In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pages 1155–1158. ACM, 2010.

[15] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding Twitter Data with Tweet Xplorer", In Proceedings of the 2013 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM,2013.

[16] Hong Yu and Vasileios Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences", In Proceedings of 8th Conference on Empirical Methods in Natural Language Processing(EMNLP'03), Sapporo, Japan, 2003.

[17] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, "Recognizing contextual polarity in phrase level sentiment analysis", In Proceedings of Human Language Technology Conference and Empirical Methods in Natural Language Processing Conference(HLT/EMNLP'05), Vancouver, Canada, 2005.

[18] Peter D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews", In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, USA, 2002.

[19] Chang, Chih-Chung, and Chih-Jen Lin, "LIBSVM: a library for support vector machines", Transactions on Intelligent Systems and Technology (TIST) 2.3: 27, ACM, 2011.

[20] O'Connor, Brendan, et al., "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", ICWSM 11: 122-129, 2010.

[21] Chang, Chih-Chung, and Chih-Jen Lin, "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3: 27, 2011.

[22] Cheng, Zhiyuan, James Caverlee, and Kyumin Lee, "You are where you tweet: a content-based approach to geo-locating twitter users", Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010.

[23] Agarwal, Apoorv, et al., "Sentiment analysis of twitter data", Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, 2011.

[24] Pak, Alexander, and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC, 2010.

[25] Greg Gorbach, "In Dynamic Market, Consumer Goods Companies Rely on Manufacturing Operations Management Systems", ARC view, 2010.

[26] Erik Cambria, Björn Schuller, Yunqing Xia and Catherine Havasi, Knowledge-Based Approaches to Concept-Level Sentiment Analysis, IEEE Transaction on Intelligent System, 2013.

[27] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford (2009): 1-12.

[28] Shamanth Kumar, Fred Morstatter, Huan Liu, "Twitter Data Analytics", Springer, 2013.

Authors' Profiles

Indrajit Mukherjee received M.Sc., Electronics degree from the University of Ranchi, India in 1995, MCA degree from BIT Mesra, Ranchi, India in 2001, PhD(Computer Science) from BIT Mesra, Ranchi, India in 2013. Currently, he is an Assistant Professor in the Department of Computer Science & Engineering, BIT Mesra Ranchi, India. His research interests include Web-Based learning, Data Mining, Big Data Handling, Web Service Applications, and Soft Computing. He has more than 15 research papers to his credit.

Jasni M Zain is Professor and Dean of Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, Malaysia. She obtained her PhD from Brunel University, West London. Her research interests include digital watermarking, image processing, data mining, cloud computing and multimedia contents. She has more than 100 research papers to her credit.

Prabhat Kumar Mahanti is Professor of Dept. of Applied Statistics (CSAS), University of New Brunswick Canada. He obtained his M.Sc. from IIT-Kharagpur, India, and Ph.D. from IIT-Bombay India. His research interests include Software engineering, software metrics, reliability modelling, modelling and simulation, numerical algorithms, finite elements, mobile and soft computing, and verification of embedded software, neural computing, data analysis and multi-agent systems. He has more than 100 research papers, technical reports to his credit.

How to cite this paper: Indrajit Mukherjee, Jasni M Zain, P. K. Mahanti, "An Automated Real-Time System for Opinion Mining using a Hybrid Approach", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.7, pp.55-64, 2016. DOI: 10.5815/ijisa.2016.07.06