A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine

D V Nagarjuna Devi[$], Chinta Kishore Kumar[#],Siriki Prasad[*]
Assistant Professor[$], UG Student[#], UG Student [*]
Department of Computer Science and Engineering,
International Institute of Information Technology,
Rajiv Gandhi University of Knowledge Technologies,
Andhra Pradesh, India
devi.duvvuri@rgukt.in[$], kishoreharishbrothers@gmail.com[#], prasadsaprs@gmail.com[*]

***Abstract*—In this modern era of globalization, e-commerce has become one of the most convenient ways to shop. Every day people buy many products through online and post their reviews about the product which they have used. These reviews play a vital role in determining how far a product has been placed in consumers' psyche. so that the manufacturer can modify the features of the product as required and on the other hand these will also help the new consumers to decide on whether to buy the product or not. However, it would be a tedious task to manually extract overall opinion out of enormous unstructured data. This problem can be addressed by an automated system called 'Sentiment Analysis and Opinion Mining' that can analyze and extract the users' perception in the whole reviews. In our work we have developed an overall process of 'Aspect or Feature based Sentiment Analysis' by using a classifier called Support Vector Machine (SVM) in a novel approach. It is proved to be one of the most effective ways to analyze and extract the overall users' view about the particular feature and whole product as well.**

***Keywords*— sentiment analysis; natural language processing; data mining; Stanford parser; product reviews.**

I. INTRODUCTION

Opinion mining is the process of detecting whether a user perception is positive or negative or neutral, whereas sentiment analysis is the process of extracting and analyzing the given data to determine the extent of positivity and negativity present in the expressed opinion. So these two terms are closely interrelated; in fact opinion mining is used in the process of sentiment analysis. It is so necessary to determine the extent of positivity or negativity in a sentence because when a user is expressing his/her views through a review, she/he may specify both the positive and negative things what s/he experienced from that product. So, a review which seems to be positive as a whole may also contain some negativity regarding some features; on the other hand there would be a possibility of having some positivity in a negative review also. Let's analyze the same points through an example. "Phone is good enough and up to the mark But camera resolution is very low", "Phone is extremely bad but the battery backup is fine"
The above thing is so evident from these two reviews. It is where sentiment analysis comes into picture.
Sentiment analysis techniques are being used in the following fields. When the government gives an opportunity to the people to express their views regarding a policy, most of the people may express their concerns about the policy. Then it would be a tedious task to analyze and understand the opinion of the people as a whole. Here this sentiment analysis can be used. It also helps in predicting the box office records of a movie based on the IMDB reviews [1]. The news forums, blogs, mail, and social media can be monitored by detecting the arrogant, vulgar, heated, and hatred words or sentences by using these techniques.

We would like to use machine learning approach for our research. The machine learning approach belongs to supervised classification approach. This approach is more accurate because every classifier is trained on a collection of data. In this approach, two types of datasets are essential. They are training and testing sets. Training set is used to train the system in such a way that it can detect the opinion expressed in the reviews accurately. Test set is used to test the performance of the classifier. Large numbers of machine learning techniques are available which classifies the data. Naïve Bayes, Maximum Entropy (ME) and Support Vector Machines (SVM) have got good results in categorization of text.

In our work we have used SVM in a novel way to find out the overall positive and negative scores for a particular feature. We finally identified overall users' perception about each and every feature and the whole product as well. We proposed a new algorithm to resolve the problem of negations.

We would like to describe about the recent trends, challenges and research in the next subsections.

978-1-4673-8286-1/16 $31.00 © 2016 IEEE
DOI 10.1109/IACC.2016.11

We will discuss our approach, evaluation of the classifier subsequently in the next sections. Finally we are going to conclude our work with our achieved results proposed future implementations.

II. RELATED WORK

A lot of research is going on in the area of sentiment analysis and opinion mining. Our work is actually motivated from the recent advancements in this machine learning techniques. We can achieve sentiment analysis in different levels as word, sentence and document levels. In many of the reviews people express diverse opinion about features. So, aspect/feature based sentiment analysis would be the most suitable for our work.

There are many machine learning approaches but we preferred SVM to classify the features. In [2]the authors have given thousand reviews to the different classifiers like Naive Bayesian, SVM, Maximum entropy, and it is proved that SVM can get the best result with high accuracy among all the other approaches.

In [3] authors proposed machine learning approaches for sentiment classification, and proved that these techniques can yield a good result when compared to other techniques.

Wang et al [4] proposed supervised learning methods. They are very popular and proved to be effective in sentiment classification. It is difficult to work with supervised methods as they are expensive and time consuming. Researches are working on this area to find out better techniques.

Opinion mining is to identify the polarity as positive or negative. Saleh, et al., [5] extracted this using Support Vector Machines (SVM) by using various datasets and weighting schemes.

After analyzing many papers we finally decided to work on support vector machine as it can the best accurate method among the all existing machine learning methods.

II. CHALLENGES IN SENTIMENT ANALYSIS

1. Sarcasm and conditional sentences: If some user uses any sarcastic expression it would be an impossible task to the system to understand its exact meaning and perception in user's point of view. Even different people can interpret a sarcastic expression in different ways, so how can we expect a system to recognize its actual intention. It is also difficult for the system to understand the conditional sentences For example:

“The phone will be awesome if its camera is good” after considering dependencies the system will interpret it as 'phone-awesome' and 'camera-good' but here the user's perception is quite different. In[6] authors proposed some semi supervised methods to address this problem.

2. Grammatical errors and poor spellings: The users may not always follow the exact grammatical rules, punctuations and spellings while writing a review, these mistakes make the system to understand the context in different way. For example 'The phone is soooooo guuuud', actually here the user wrote ‘soooooo’, ‘guuuud’ instead of ‘so’ and ‘good’ to show his deeper feeling, but the system will interpret ‘soooooo’ and ‘guuuud’ as nouns though they are adverb, adjective respectively, in the user's point of view. In [7] Dey and Haque, proposed some approaches to rectify these errors.

3. Spam Detection: Some users’ usually try to post the negative reviews to spoil others' reputation. Nowadays it has become the biggest challenge to detect the spam among tons of reviews in the internet. So, here is a dire need to develop a system that can detect and remove the spam. In [8] Bing liu proposed some methods to detect spam in the reviews.

4. Anaphora Resolution: Most of the researchers are ignoring the pronouns during sentiment analysis as it is so ambiguous to the system to detect what a pronoun or noun phrase is actually referring to in the sentence, but sometimes even a pronoun may also refer something which is so vital in extracting users' intention. For Example...'I love the phone. It is quite good'

Here even if the word 'It' is acting as a pronoun in the second sentence, it has some importance as it is referring 'phone' in the first sentence. Without understanding that reference we cannot correlate the opinion word 'good' to the 'phone', which the user is actually specifying according to the context. In [9] the authors discussed a novel approach to resolve the problem of reference by using some appreciable methods.

III. THE PROPOSED METHOD

A. Dataset

Dataset is the collection of all reviews which we will give as input to the system. There are 2 types of datasets 1) Training set 2) Testing set. Training sets are the reviews which we take from some e-commerce sites so as to train the system. After this training, system can able to understand which things undergo positive and which things fall under negative for a particular feature of a product. After completion of the training we will give set of input reviews whose polarities are already known to us. Based on the output given by the system we can understand the accuracy of the system and can proceed further. These reviews

come under testing reviews. Here we are taking collection of reviews of different laptop companies like HP, APPLE, DELL, LENOVO etc. as datasets and reviews from e-commerce sites like amazon, eBay, flip kart, etc. as training sets and finally some specific reviews whose polarities are known to us are taken as test set.

B. Sentence Level Classification

There would be 3 types of opinions that can be expressed through any sentence in the world they are positive, negative, neutral in a review user may write positively or negatively about a particular feature, those are needed to be considered as they are specifying some polarity. They are called subjective sentences. But sometimes user may use some normal sentences where he doesn't specify his opinions or simply asks a question in the middle of a review. Those sentences will come under objective sentences. We need to remove all of them so that the future procedure should not be messy. The presence of relative pronouns like where, whom, who, how etc. may increase the probability of sentence being a question. So, we have to remove those sentences by detecting using POS tagging. There will be huge opinionated words in the SentiWordNet with some polarity. If the words present in the reviews not matched with the any opinionated word from the list we have to remove that from the review.

C. Extraction of Aspects

It's the most important & challenging thing in the sentiment analysis. The aspect/feature of any particular product would be most of the times a noun or noun phrase. We need to identify the features during the time of training phrase itself. For that we can use POS tagging and can detect words with tags like NNS (noun plural), NN (Noun), NNP (proper noun singular) etc. Minimum support threshold is used to find all the features about which the users expressing their views frequently. The infrequent features are to be discarded. For a product like laptop the frequent features would be battery, screen, processing speed, storage etc.

E. Extraction of Opinion Words for Aspects

Opinion word extraction for an aspect can be achieved by using Stanford parser. Parser is a tool that can find the grammatical dependencies between the words in the sentences. When we give review as input to the parser it will produce dependency sets as the output. We need to observe dependencies to find the opinion word for the aspects which we have already extracted in the previous step. There will be many grammatical dependencies among them, probably the following contain opinion words along with aspects they are amod(adjectival modifier),

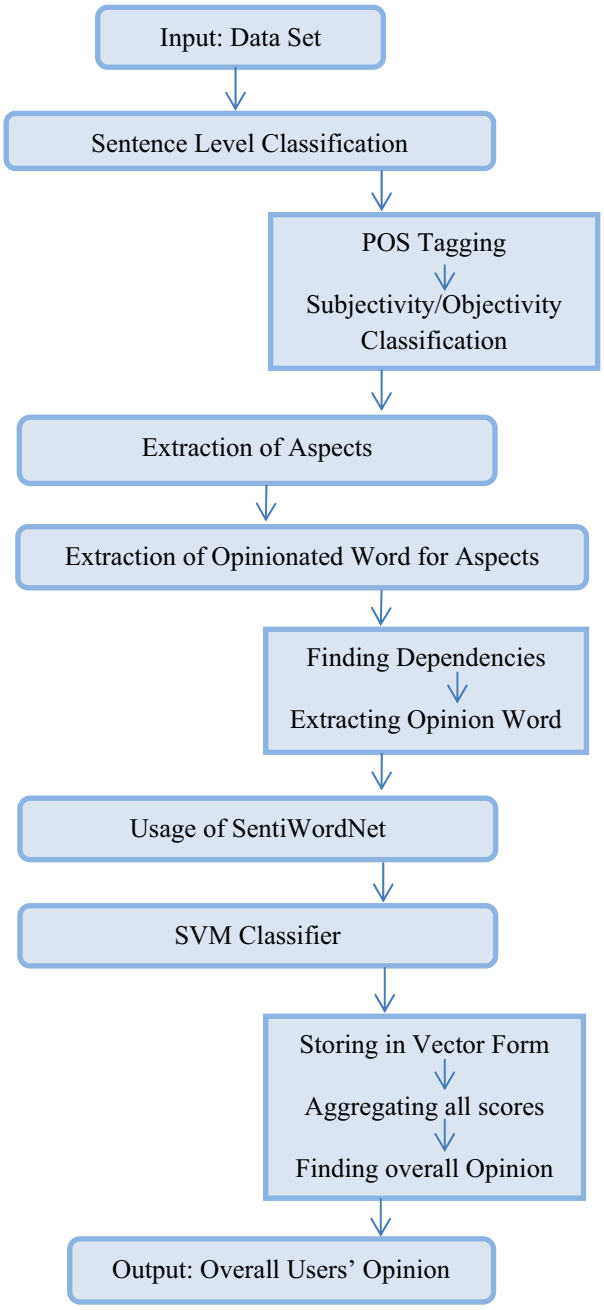


Figure1: Overall Procedure Diagram

xcomp(clausal component with external subject), advmod(adverbial modifier),nsubj(nominal subject) and negation modifier. Whenever we are observing the dependencies sometimes we can directly find the opinion words for some aspects such type of dependencies are called direct dependencies. Even then we may not extract all the opinion words then we need to check for the transitivity among the dependencies to extract the remaining opinion words. Now the remaining dependencies can be discarded.

Let us illustrate the same thing with an example.
"I love the laptop as i am satisfied with the speed, but cost is high."

Tagging:
I/PRP love/VBP the/DT laptop/NN as/IN i/FW am/VBP satisfied/VBN with/IN the/DT speed/NN/, but/CC cost/NN is/VBZ high/JJ ./.

Universal Dependencies:
nsubj(love-2, I-1)
root(ROOT-0, love-2)
det(laptop-4, the-3)
nsubjpass(satisfied-8, laptop-4)
case(i-6, as-5)
nmod(laptop-4, i-6)
auxpass(satisfied-8, am-7)
ccomp(love-2, satisfied-8)
case(speed-11, with-9)
det(speed-11, the-10)
nmod(satisfied-8, speed-11)
cc(love-2, but-13)
nsubj(high-16, cost-14)
cop(high-16, is-15)
conj(love-2, high-16)

From the above dependencies it is clear that nsubj(high-16, cost-14), nmod(satisfied-8, speed-11) are direct dependencies. Here, for the aspects 'cost' and 'speed' the opinion words 'high' and 'satisfied' respectively are derived from direct dependencies. But opinion word 'satisfied' for the aspect 'laptop' is extracted through transitivity of nsubjpass(satisfied-8, laptop-4) and ccomp(love-2, satisfied-8). In the same way, all the sentences can be parsed and a group of opinion words for each aspect from various sentences can be collected. Whenever we find a negation in either direct or transitive relation, the opinion word will be joined with minus symbol. This will be used for negation representation in the future steps.

F. SentiWordNet

SentiWordNet is a tool specially designed for sentiment analysis application. In this, every word is associated with two types of polarities namely positive and negative. The scores will be different for the two cases for a specific word. For example,

"Cost is high"

actually the word high is positive in nature but in this context the cost is high means the customer couldn't able to afford it, In turn it's conveying a negative sense from this it is clear that the same word is showing different polarities in different contexts, so during training process we will train the system in such a way that it can recognize the polarity of the sentence based on the given context.

F. Support Vector Machine

We have chosen SVM as the classifier in our work. This technique was introduced by[10], and used for two group classification. Each group consists of sets of vectors. Then every data represented as vector is classified in a particular class. The unique property of SVM is that it can learn even if we provide huge data. SVM works well for text classification because it can handle large features. Another advantage of SVM is that it is robust when there is a small set of examples distributed over a large area and also because most of the data sets of linearly separable. SVM has given reliable results in the past research in sentiment analysis [2].

SVM can classify by the cases into different categories by constructing hyper planes. SVM classifier should first be trained by using training set reviews so that classifier can learn effective categorization. The training of the system can be done by using the feature set $x_1 : y_1$, $x_2 : y_2$, ..., $x_m : y_m$. After the training process, machine learns a classifier. Here x is a product feature which is extracted by using Stanford parser and y is the score of that opinion word which extracted by using SentiWordNet. By using these training sets we can define hyper plane that effectively separates two classes.

The result of the classifier is a set of vectors that contains the aspect and its opinion words. The classifier output is stored in the form of following vector,

$$V_k(R_m,F_i,O_{ij},P_{ij},N_{ij})$$

Where, V_k is k^{th} vector, R_m is m^{th} review, F_i is i^{th} feature in review m, O_{ij} is j^{th} opinionated word for i^{th} feature, P_{ij} is Positive score for j^{th} opinion word of i^{th}feature, N_{ij} is Negative score for j^{th} opinion word of i^{th} feature.

Extraction of feature wise opinion:

In order to determine the users overall perception about a particular feature, at first we considered all the vectors with that feature (f_j). Let's say there are z vectors.

$$\text{Final positive score for particular feature} = \frac{\sum_{i=1}^{z} p_i}{z}$$

$$\text{Final negative score for particular feature} = \frac{\sum_{i=1}^{z} n_i}{z}$$

In the above two formulas numerators represent the sum of all the positive and negative

scores for j^{th} feature respectively. Denominators represent number of vectors considered.

Now among all those z vectors we separated vectors with same review number and added the positive and negative scores separately.

For every R_i=k, let's say we have n such vectors. Where k is a particular review number

Sum of positive scores = $\sum_{i=1}^{n} p_i$

Sum of negative scores= $\sum_{i=1}^{n} n_i$

Now store these scores in vectors along with their review number. Finally we have ended up with vectors with unique review number for a particular feature. Let's take these vectors as m, we count the number of vectors in m whose positive score is greater than negative score that count would be the number of positive reviews about that particular feature and the rest of the vectors are negative ones about that feature.

Positive opinioned reviews=No of vectors in m whose positive score is greater than negative score.
Negative opinioned reviews=m-positive opinioned reviews.

Extraction of Whole product opinion:

To find the users' opinion about a whole product, we considered all the vectors and among them for every particular review numbered vectors we sum up their positive and negative scores separately. Now we have ended up with only one positive and negative score for each review. We count the number of vectors whose positive score is greater than negative score that count would be the number of positive reviews about that product and the rest of the vectors are negative ones about that product.

If we get any opinionated word with -1 as output of classifier we know that negation (positive)=negative and negation(negative)=positive so while storing the scores in the vectors will interchange the positive and negative scores. It will automatically resolve the problem of usage of negation in the sentences.

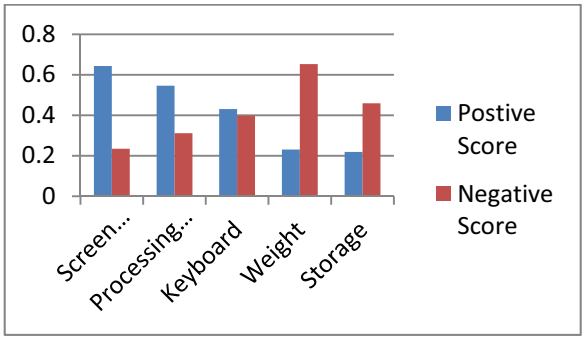
G. Results

We have given the test set which consists of hundred reviews to system whose positive and negative reviews count of different features already known to us, then it produces the number of positive and negative reviews of different features of a particular product as output. These values are depicted below as table1.

	Practical Values		Actual Values	
	+ve	-ve	+ve	-ve
Screen Resolution	63	12	70	16
Processing speed	56	20	50	26
Keyboard	26	21	30	25
Weight	21	54	16	58
Storage	12	45	17	51
Overall opinion	**76**	**24**	**79**	**21**

Table1: Actual and Practical number of positive and negative reviews

Overall positive and negative scores of particular feature are also calculated and are depicted in the graph1 below.



Graph1: Overall scores of each feature

Based on the actual and practical number of positive and negative reviews we will calculate the accuracy of the classifier.

G. Evaluation

During the evaluation of performance of the classifier we have considered four parameters as depicted in the table2 below. By using these four parameters Evaluation Measures like Accuracy, precision and F-measure can be calculated.

		Correct Labels	
		Positive	Negative
Classified Labels	Positive	TP(True Positive)	FP(False Negative)
	Negative	FN(False Negative)	TN(True Negative)

Table2: Contingency Table

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

After calculating the values of precision, recall, accuracy these values are shown in the below table and also overall accuracy of the SVM is calculated.

	Precision	Recall	Accuracy
Laptop1	91.04%	89.43%	90.85%
Laptop2	83.78%	89.55%	85.21%
Laptop3	86.64%	90.32%	88.34%
Final	87.15%	89.76%	88.13%

Table3: Overall accuracy of SVM

We can get more accuracy by using SVM when we have few number of features [11]. There are some other factors which can affect the efficiency like domain and data set size [5].

H. Conclusion and Future Work

In this manuscript we have proposed a new way of using SVM as a classifier and it is proved to be an effective method to find users' perception about a feature and product also. We proposed a novel way of resolving the problem of negation that usually appears in any review.

Future research can focus on sarcastic expressions which are usually difficult to understand, both by the users' and the computer system. One more challenging issue is the detection of spam contents in users' review. Finally the study can be extended to resolve the problem of co-reference resolution.

REFERENCES

1. Pruthvi H.R, Nagamma P, Shwetha N H and Nisha K.K*An Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction,* In the proceedings of International Conference on Computing, Communication and Automation (ICCCA2015)

2. Bo Pang, Shivakumar Vaithyanathan ,Lillian Lee. *Thumbs up? Sentiment classification using machine learning techniques.* In Conference on Empirical Methods in Natural Language Processing. 2002

3. Pang B & Lee L. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*, The Association for computational linguistics, pp. 271–278, 2004.

4. Wang Z, Li S, Lee S.Y.M, Zhou G, '*Semi-supervised learning for imbalanced sentiment classification*', international joint conference on artificial intelligence, pp. 1826–1831, 2012.

5. Martín-Valdivia M T, Rushdi Saleh M, Ureña-López L A, Montejo-Ráez A, *Experiments with SVM to classify opinions in different domains*, Expert Systems with Applications, 38(12), 14799-14804, 2011.

6. Ari Rappoport, Oren, Tsur, Dmitry Davidov, *A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews*, Fourth International AAAI Conference on Weblogs and Social Media 2010.

7. S K Mirajul Haque and Dey, Lipika, *Opinion mining from noisy text data*, Second Workshop on Analytics for Noisy Unstructured Text Data 2008.

8.Bing Liu, *Sentiment Analysis and Subjectivity*, *Handbook of Natural Language Processing*, 2010.

9. Mihai Surdeanu,, Yves Peirsman, Nathanael Chambers, and Dan Jurafsky, Angel Chang, Heeyoung Lee. *Deterministiccoreference resolution based on entity-centric, precision rankedrules*. In the proceedings of Computational Linguistics, 2013.

10. V. Vapnik and C. Cortes, *Support-vector networks,* Handbook of Machine Learning, 1995.

11. X. Liu, Y. Shi ,E. Haddi, *The role of text pre-processing in sentiment analysis*, International Conference on Information Technology and Quantitative Management 2013.