Implementation of Support Vector Machine Technique in Feedback Analysis System

Sheetal Pereira
K.J.Somaiya COE
Vidyavihar,Mumbai

Uday Joshi
K.J.Somaiya COE
Vidyavihar,Mumbai

ABSTRACT

E-commerce is very popular and interactive these days. Industries that produce new products and selling them on the Web often ask their customers to review the products which they have purchased. These reviews help both i.e. customers as well as producers. Producers get the idea from reviews about how their customers feel about the product which they have purchased and customers who want purchase the product read these reviews and take decision. As the number of responses are available very large in number, manually organizing large set of reviews/responses into required categories and analyzing them is time consuming, expensive and is often not feasible. So automated text classification is done to overcome these constraints. Various techniques can be used for automatic text classification. In this paper Support Vector Machine (SVM) which is a supervised learning technique is used in feedback analysis system which accepts the responses given by students as input preprocess it and lastly applies term weighting algorithm. After applying term weighting algorithm it displays analysis to the particular faculty.

General Terms

Web, E-commerce, Responses, Techniques

Keywords

Support Vector Machine (SVM), Supervised Learning Technique

1. INTRODUCTION

The feedback analysis system takes the feedback from students, analyzes it and displays the analysis to the faculty. The number of responses that are available is very huge in number. It is difficult for the faculty to go through the responses given by students. So Support Vector Machine (SVM) is used for analyzing these responses. Here Support Vector Machine (SVM) is not used to categorize theses responses to its predefined categories but it is used in analysis [1]. As the predefined category to which these responses fall under is only one that is educational.SVM is a supervised machine learning approach [2], it can be used as a binary text classifier which classifies responses into two categories viz. positive and negative [3] and accordingly the responses are analyzed. In this paper, SVM is used as a binary text classifier [4].

SVM is used in many applications such as document organization, text filtering and hierarchical categorization of web pages. Document organization means the task of structuring documents of a corporate document base into folders [5]. Text filtering is the process of classifying a dynamic collection of text document into two disjoint categories such as relevant and irrelevant [6]. Similarly, hierarchical categorization of web pages helps users in search and browse operations by posing a generic query in the hierarchy of categories and restricting the search to the particular categories of interest rather than posing to a general purpose search engine [7].

2. RELATED WORK

There are various approaches to text classification viz. Naïve Bayes, Maximum Entropy, SVM [8]. Naïve Bayes is a probabilistic learning method which applies Bayes theorem [9]. Maximum Entropy is a probability distribution estimation technique, the principle is without external knowledge distribution should be uniform [10].

Support Vector Machine (SVM) is the only classifier which can be implemented as a binary text classifier i.e.it classifies the responses into two categories viz. positive and negative. There are various implementation steps:

- Parsing the documents and case-folding.
- Removing stopwords
- Stemming
- Dimensionality reduction
- Term weighting

These steps are described in detail in the following sections. The first three steps are called preprocessing steps as they remove the unwanted data.

2.1 Parsing the documents and case-folding

This step removes all HTML tags and non-alpha characters from the document. Case-folding means converting all the characters in a document into the same case. In this paper all characters are converted into lower-case. In this step tokens consisting of alpha characters are extracted. This step is also called as tokenization.

2.2 Removing stopwords

There are words in English which are used to provide structure to the language like conjunctions, articles, pronouns and prepositions. Such words which occur very frequently and carry no useful information about the content are called stopwords. So, remove such words from the document. After removing stopwords, document will remain with the important words that mean the words which are important for text classification.

2.3 Stemming

Stemming is the process for reducing derived words to their stem or root. In this step the words which are in the same context with the same term are reduced to their root. For example, we reduce the similar terms “teaches”,” taught”, and “teaching” to the stem word “teach”. Porter’s stemming

Algorithm is applied [11]. Porter stemmer utilizes suffix stripping.

Porter's stemming algorithm steps:

1. Gets rid of plurals and –ed or –ing suffixes.
2. Turns terminal y to i when another vowel in stem.
3. Maps double suffixes to single ones:-ization, -ational etc.
4. Deals with suffixes, -full, -ness etc.
5. Takes off –ant, -ence, etc.
6. Removes a final –e.

First three steps are called preprocessing step. After preprocessing step unwanted data is removed from the document. That means document remains with the data which is important for analysis. Next Part-Of-Speech (POS) tagger is applied to the document. POS tagger is software which reads the text in some language and assigns parts of speech to each word such as verb, noun, adjective etc.[12].Using POS tagger positive and negative words are identified and if it matches with the word in the database counter is incremented.

2.4 Dimensionality Reduction

This step is also called as Feature Selection. In this step the features/dimensions which are irrelevant for the analysis are deleted. In this paper, as all the features are important, this step is not considered.

2.5 Term Weighting

In this step, weight is assigned to a word based on number of times it occurs in the document. This method is called term frequency and inverse term frequency which is a traditional method to assign a weight to the words. In this paper, the number of positive and negative words is counted and accordingly weights are calculated. Following formula is used to calculate weight.

Count=2*positive count - 1*negative count

Depending on the count value message is displayed to the faculty.

3. DISCUSSION

In this section advantages and disadvantages of Support Vector Machine (SVM) and Maximum Entropy (ME) are discussed below in the given Table 1:

Table 1: Advantages and disadvantages of SVM and ME techniques

Technique	Advantages	Disadvantages
SVM	It is highly accurate	Its speed is low
	It can handle many features	It requires more time to process
ME	Its speed is fast	It is less accurate
	It requires less time to process	Its efficiency is low

4. EXPERIMENTAL RESULTS

This section represents the implementation results of Support Vector machine technique for feedback analysis system. All experiments are performed on Intel i5-480M processor with 4 GB memory using SQL server 2005 as a backend for database and JSP and Java for front end programming. A huge database of approximately 500 records of students and 500 records of responses are used. User login screens are created. If the user is a student then it will open a feedback page but if the user is a faculty it will ask that user to choose the subject he/she has taught and accordingly it will display the analysis of feedback using SVM to the faculty.

4.1 The algorithm for feedback analysis using SVM

Step 1: Registered users login into the system using username and password, it checks the credentials of the user with the database whether the user is a student or faculty else user has to register with the system.

Step 2: If the user is a student it will display the feedback form to the student.

Step 3: But if the user is a faculty it will ask faculty to choose the subject which he has taught.

Step 4: The chosen subject's id and faculty's id are matched with the collected feedback and accordingly feedback are separated and stored in a text file. At this stage each faculty's feedback is separated and kept it in a text file.

Step 5: This text file is given as an input to the tokenizer which removes non-alpha characters from the document and converts all the characters to the same case. And again the result is stored in another text file.

Step 6: This stage takes the text file as an input given by the previous step and applies stop word removing algorithm to it and stores the result in another text file. In this stage the unwanted words are removed.

Step 7: This stage takes the text file as an input given by the previous step and applies stemming algorithm to it and stores the result in another text file.

Step 8: At this stage the document contains only the important words. In this step POS tagger is applied to the text file which assigns part-of-speech to each word such as verb, adjective, noun etc. and stores the result in another text file.

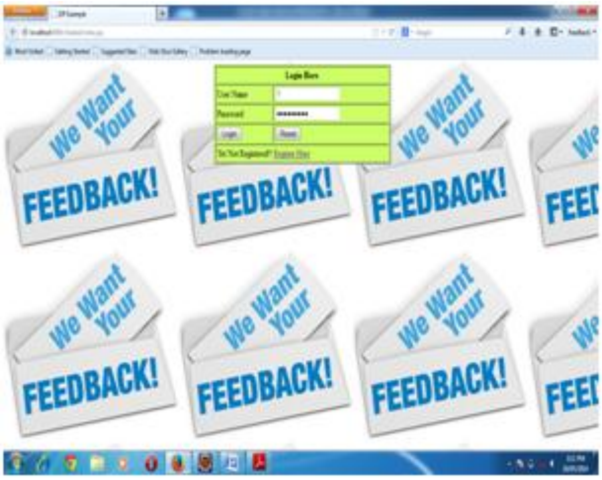
Step 9: This step takes the text file as an input given by the previous step and matches positive and negative words with the database which contains all positive and negative words and accordingly increments counter for positive and negative words.

Step 10: In this step count value is calculated based on the following formula:

Count= 2* positive count – 1* negative count

Depending on the count value message is displayed to the faculty. For example if numbers of responses given are 10 then maximum value of count for positive responses will be 20 and minimum value of count will be -10.If count is 20 then teaching is excellent and if it is 10 then teaching is good like this message is displayed to the faculty.

First the user has to login into the system using username and password. Output in figure 1 shows the login screen for the user.

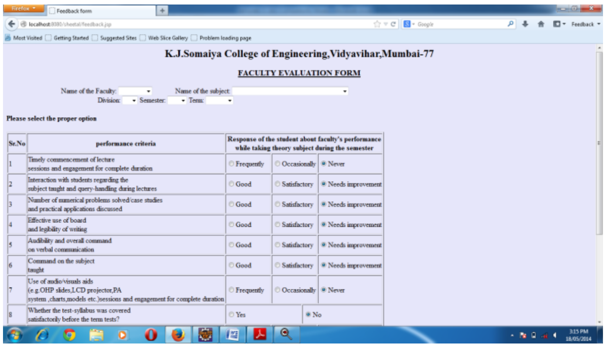


Login Box

User Name	
Password	●●●●●●●●
Login	Reset
Its Not Registered? Register Here	

Figure 1: Login screen for the user

If the user is a student then it will display the feedback form to the student. Figure 2 shows the feedback form for students.



K.J.Somaiya College of Engineering,Vidyavihar,Mumbai-77

FACULTY EVALUATION FORM

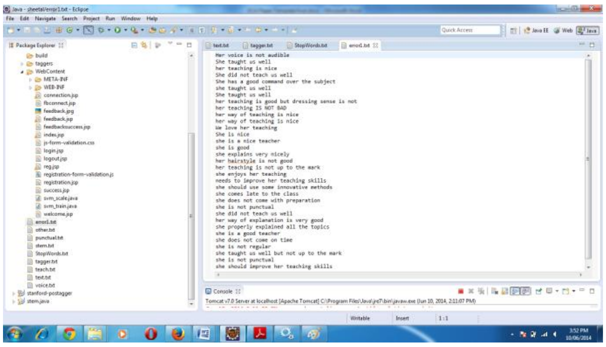
Name of the Faculty: Name of the subject:
Division: Semester: Term:

Please select the proper option

Sr.No	performance criteria	Response of the student about faculty's performance while taking theory subject during the semester		
1	Timely commencement of lecture sessions and engagement for complete duration	○ Frequently	○ Occasionally	● Never
2	Interaction with students regarding the subject taught and query-handling during lectures	○ Good	○ Satisfactory	● Needs improvement
3	Number of numerical problems solved/case studies and practical applications discussed	○ Good	○ Satisfactory	● Needs improvement
4	Effective use of board and legibility of writing	○ Good	○ Satisfactory	● Needs improvement
5	Audibility and overall command on verbal communication	○ Good	○ Satisfactory	● Needs improvement
6	Command on the subject taught	○ Good	○ Satisfactory	● Needs improvement
7	Use of audio/visuals aids (e.g.OHP slides,LCD projector,PA system ,charts,models etc.)sessions and engagement for complete duration	○ Frequently	○ Occasionally	● Never
8	Whether the test-syllabus was covered satisfactorily before the term tests?	○ Yes	● No	

Figure 2: Feedback form for students

If the user is a faculty then it will ask him to choose the subject taught. Once the subject is chosen it will separate all responses in the separate text file according to faculties. Figure 3 shows responses for a particular faculty.

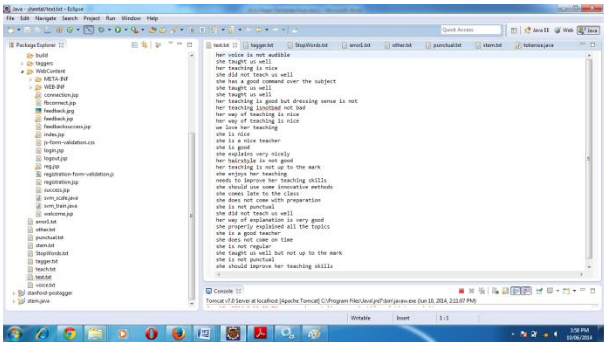


```
Her voice is not audible
She taught us well
her teaching is nice
She did not teach us well
She has a good command over the subject
she taught us well
She taught us well
her teaching is good but dressing sense is not
her teaching IS NOT BAD
her way of teaching is nice
her way of teaching is nice
We love her teaching
She is nice
she is a nice teacher
she is good
she explains very nicely
her hairstyle is not good
her teaching is not up to the mark
she enjoys her teaching
needs to improve her teaching skills
she should use some innovative methods
she comes late to the class
she does not come with preparation
she is not punctual
she did not teach us well
her way of explanation is very good
she properly explained all the topics
she is a good teacher
she does not come on time
she is not regular
she taught us well but not up to the mark
she is not punctual
she should improve her teaching skills
```

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre7\bin\javaw.exe (Jun 10, 2014, 2:11:07 PM)
```

Figure 3: Responses of a particular faculty

Next, these responses are tokenized that is non- alpha characters are removed and all characters are converted to the same case. Figure 4 shows the result after tokenization.

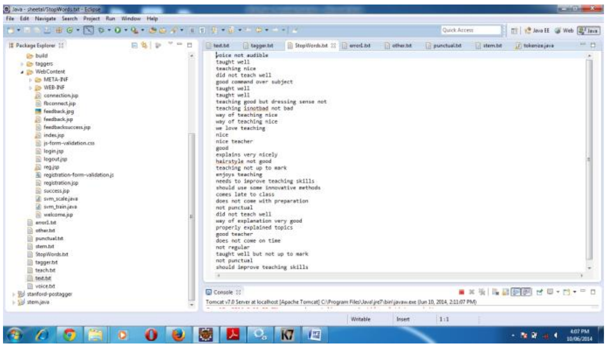


```
her voice is not audible
she taught us well
her teaching is nice
she did not teach us well
she has a good command over the subject
she taught us well
she taught us well
her teaching is good but dressing sense is not
her teaching isnotbad not bad
her way of teaching is nice
her way of teaching is nice
we love her teaching
she is nice
she is a nice teacher
she is good
she explains very nicely
her hairstyle is not good
her teaching is not up to the mark
she enjoys her teaching
needs to improve her teaching skills
she should use some innovative methods
she comes late to the class
she does not come with preparation
she is not punctual
she did not teach us well
her way of explanation is very good
she properly explained all the topics
she is a good teacher
she does not come on time
she is not regular
she taught us well but not up to the mark
she is not punctual
she should improve her teaching skills
```

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre7\bin\javaw.exe (Jun 10, 2014, 2:11:07 PM)
```

Figure 4: Result after tokenization

Next step is to remove stop word that is removing the words which are not important from analysis point of view. So the words which give only structure to the language are removed. Figure 5 shows the result after stop word removal algorithm.

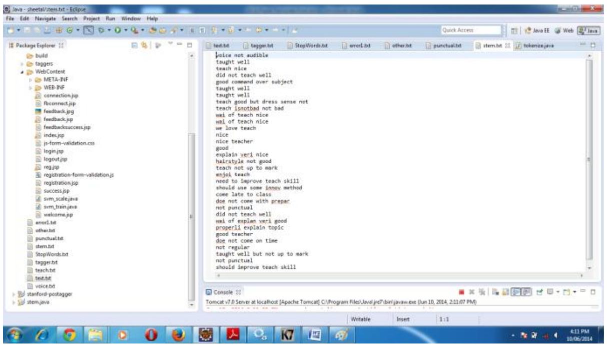


```
voice not audible
taught well
teaching nice
did not teach well
good command over subject
taught well
taught well
teaching good but dressing sense not
teaching isnotbad not bad
way of teaching nice
way of teaching nice
we love teaching
nice
nice teacher
good
explains very nicely
hairstyle not good
teaching not up to mark
enjoys teaching
needs to improve teaching skills
should use some innovative methods
comes late to class
does not come with preparation
not punctual
did not teach well
way of explanation very good
properly explained topics
good teacher
does not come on time
not regular
taught well but not up to mark
not punctual
should improve teaching skills
```

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre7\bin\javaw.exe (Jun 10, 2014, 2:11:07 PM)
```

Figure 5: Result after stop word removal

After removing stop words next step is to apply Porter's stemming algorithm and reduce the words which are there in the same context to their stem. Figure 6 shows the result after Porter's stemming algorithm.

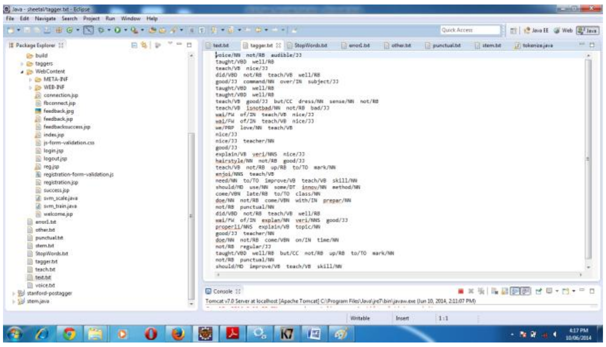


```
voice not audible
taught well
teach nice
did not teach well
good command over subject
taught well
taught well
teach good but dress sense not
teach isnotbad not bad
wai of teach nice
wai of teach nice
we love teach
nice
nice teacher
good
explain veri nice
hairstyle not good
teach not up to mark
enjoi teach
need to improve teach skill
should use some innov method
come late to class
doe not come with prepar
not punctual
did not teach well
wai of explan veri good
properli explain topic
good teacher
doe not come on time
not regular
taught well but not up to mark
not punctual
should improve teach skill
```

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre7\bin\javaw.exe (Jun 10, 2014, 2:11:07 PM)
```

Figure 6: Result after Porter's stemming algorithm

After stemming algorithm all unwanted words are removed and the document remains with only important words. Next POS tagger is applied which assigns part-of-speech to each word. Figure 7 shows the result after POS tagger.

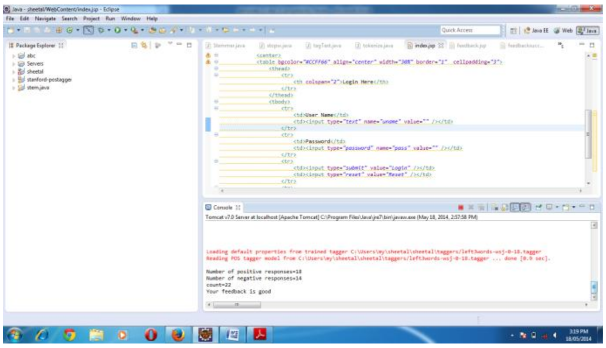


```
voice/NN  not/RB  audible/JJ
taught/VBD  well/RB
teach/VB  nice/JJ
did/VBD  not/RB  teach/VB  well/RB
good/JJ  command/NN  over/IN  subject/JJ
taught/VBD  well/RB
taught/VBD  well/RB
teach/VB  good/JJ  but/CC  dress/NN  sense/NN  not/RB
teach/VB  isnotbad/NN  not/RB  bad/JJ
wai/FW  of/IN  teach/VB  nice/JJ
wai/FW  of/IN  teach/VB  nice/JJ
we/PRP  love/NN  teach/VB
nice/JJ
nice/JJ  teacher/NN
good/JJ
explain/VB  veri/NNS  nice/JJ
hairstyle/NN  not/RB  good/JJ
teach/VB  not/RB  up/RB  to/TO  mark/NN
enjoi/NNS  teach/VB
need/NN  to/TO  improve/VB  teach/VB  skill/NN
should/MD  use/NN  some/DT  innov/NN  method/NN
come/VBN  late/RB  to/TO  class/NN
doe/NN  not/RB  come/VBN  with/IN  prepar/NN
not/RB  punctual/NN
did/VBD  not/RB  teach/VB  well/RB
wai/FW  of/IN  explan/NN  veri/NNS  good/JJ
properli/NNS  explain/VB  topic/NN
good/JJ  teacher/NN
doe/NN  not/RB  come/VBN  on/IN  time/NN
not/RB  regular/JJ
taught/VBD  well/RB  but/CC  not/RB  up/RB  to/TO  mark/NN
not/RB  punctual/NN
should/MD  improve/VB  teach/VB  skill/NN
```

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre7\bin\javaw.exe (Jun 10, 2014, 2:11:07 PM)
```

Figure 7: Result after POS tagger

After applying POS tagger in the next step, positive and negative words is counted and variable count is calculated and depending on the count value the message is displayed to the faculty. Figure 8 shows the analysis of the feedback/responses to the faculty.



```
<center>
<table bgcolor="#CCFF66" align="center" width="308" border="1"  cellpadding="3">
    <thead>
        <tr>
            <th colspan="2">Login Here</th>
        </tr>
    </thead>
    <tbody>
        <tr>
            <td>User Name</td>
            <td><input type="text" name="uname" value="" /></td>
        </tr>
        <tr>
            <td>Password</td>
            <td><input type="password" name="pass" value="" /></td>
        </tr>
        <tr>
            <td><input type="submit" value="Login" /></td>
            <td><input type="reset" value="Reset" /></td>
        </tr>
```

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre7\bin\javaw.exe (May 18, 2014, 2:57:58 PM)

Loading default properties from trained tagger C:\Users\my\sheetal\sheetal\taggers/left3words-wsj-0-18.tagger
Reading POS tagger model from C:\Users\my\sheetal\sheetal\taggers/left3words-wsj-0-18.tagger ... done [0.9 sec].

Number of positive responses=18
Number of negative responses=14
count=22
Your feedback is good
```

Figure 8: Analysis shown to faculty

As the number of positive responses is more it displays “Your feedback is Good”.

5. CONCLUSION

In this paper support vector machine which is a supervised machine learning technique is used for text classification. Feedback analysis system accepts the responses from students’, processes those responses and displays the analysis to faculties. Support Vector Machine technique was successfully implemented for feedback analysis system on SQL server/JSP/Java. The algorithm was tested for approximately 500 records considered as a dataset. Implementation of SVM gives the analysis of responses to the faculty. The output of this implementation can be further used for analyzing responses using hybrid machine learning techniques which further applies K-means algorithm to it and analyzes the responses according to different topics like teaching, punctuality, audibility etc. and gives topic wise analysis the faculty.

6. REFERENCES

[1] Anand Mahendran, Anjali Duraiswamy, Amulya Reddy, Clayton Gonsalves,” Opinion mining for Text Classification”, Tech., Inst., Cognition and Learning, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume No.2, Issue No.6, pp : 589-594 ,June 1,2013.

[2] Thorsten Joachims,” text Categorization with Support Vector Machines: Learning with many relevant features” University Dortmund, Germany, 1998.

[3] Larkey ,L.S,”A patent search and classification system ”,Proceedings of 4th ACM Conference on Digital Libraries,pp.179-187,Berkely,US,1999.

[4] Istvan Pilaszy,”Text categorization and Support Vector Machines”,2005

[5] Sebastiani ,F., ”Machine learning in automated text categorization”, ACM Computing Surveys,Vol.34, No.1,pp.1-47,2002.

[6] Koller, D. and M.Sahami, “Hierarchically classifying documents using very few words”, Proceedings of 14th International Conference on machine Learning,pp.170-178,Nasshville, US,1997.

[7] Dumais , S.T. and H.Chen, “Hierarchical classification of web content ” Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, pp. 256-263,Athens,GR ,2000.

[8] Bo pang and Lillian Lee and Shivakumar Vaithyanathan,”Thumps up? Sentiment Classification using machine Learning Techniques”, Cornell university,2002.

[9] S.L.Ting,W.H.Ip,Albert H.C. Tsang “Is Naïve Bayes a Good Classifier for Document Classification?” International Journal of Software Engineering and Its Applications Vol.5, No.3,July 2011.

[10] Kalam Nigam, John Lafferty, Andrew McCallum,” Using Maximum Entropy for text classification” School of computer science Carnegie Mellon University Pittsburgh,2002.

[11] M.F.Porter, "An algorithm for suffix stripping", Computer laboratory, corn exchange street, Cambridge, 2006

[12] Nlp.stanford.edu/software/tagger.shtml