THE INTERNATIONAL JOURNAL OF
SCIENCE & TECHNOLEDGE

Sentiment Analysis and Effective Visualization of Faculty and Course Feedback

Divya R.
Student, Department of CSE, School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India
Sandhya S.
Student, Department of CSE, School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India
Vishnu Sai S.
Student, Department of CSE, School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India

Abstract:
One of the most important ways to ensure the success of any product is to understand the mindset of the users/consumers. In order to achieve that understanding, one has to collect their opinion in the form of feedback. Hence feedback collection serves as an important goal towards the promotion of the product. This can be analogous to any educational body like a School or a University. The standard of the education can be improved greatly through the feedback data obtained from the students. But the key point here is to correctly interpret the opinions collected. Sentiment analysis techniques, as the name implies aids in understanding the sentiments of the users/students. These techniques determine whether the sentiment has a penchant towards positivity or negativity majorly. Although there is the concept of neutrality, this paper aims on categorizing the comments as positive and negative. The main aim of this paper to analyze the feedback data collected to come up with results that focus on the highlighting how much the students are learning and improving themselves and how much the faculty and the courses offered are helping them achieve that. Students' opinions on both faculty and course are collected in the form of comments through an online form. The data collected is then subjected to Cleaning and Tagging. The parameters that are taken into account for the faculty are communication skills, knowledge base, motivating factor and practical implementation of concepts. There is also an extra column for students to enter other opinions if any. Similarly, for the course feedback the parameters considered are relevance of course, course coverage, availability of learning material and scope for learning new concepts. The opinions collected are then assigned a specific value called Semantic Orientation (SO) based on the polarity of the comment, i.e. whether the comment falls under positive or negative category. Analysis is then performed based on the SO value to get an idea about the polarity of the comment. The values are then plotted as analytical graphs as an attempt to provide better visualization of the obtained results. The proposed system does better analysis at that it does not rely totally on scaling to obtain the results.

Keywords: *Sentiment analysis, feedback, positive, negative, visualization, polarity*

1. Introduction

The 21st century is an age which can be considered as a digital library filled with information and opinions. Data is the new currency in this age and every corporation is very interested in getting a public opinion of their products and services. The companies that receive proper feedback tend to learn a lot better than others on the take of users' mentality and requirements. So the need of the hour is not just receiving good feedback from customers but also performing analysis on the feedback available to extract useful information that helps improve the standard of the organization. This is because new customers tend to consult the feedback provided by existing ones before making a decision. Sentiment analysis or Opinion mining is the field that has been attracting a large number of researchers owing to the availability of huge volumes of data from various sources. The data available is so unstructured that in some cases the analysis of this unstructured text becomes even more vital than extraction of useful information.

Sentiment analysis involves a lot of pre-processing before the actual analysis. This is necessary because most of the data is unstructured with a lot of slang words (wth, btw), emoticons (:-), : P), short words (gud, gr8) etc. Along with these, hyperlinks and images may also be included in the reviews. So it is vital to extract the relevant data from this unstructured data and remove any unnecessary words. After the data is cleaned, it is analyzed to determine the polarity or orientation of the reviews and comments. There are various ways to determine the polarity of comments. They can be primarily categorized into, a Lexicon based approach and a Learning based approach. The Lexicon based approach utilizes a predefined corpus of words to determine the polarity of opinions. The Learning based approach on the other hand, uses sentiment classifiers to do the same. We are using the SO-PMI-IR learning method in our project to calculate the semantic orientation of the feedback data and categorize them into positive and negative.

2. Literature Survey

The feedback collected on the faculty and course of the university is initially subjected to preprocessing techniques, mainly data cleaning. The irrelevant information like punctuation marks, smiley and the like are removed. The cleaned text is then tagged using a POS tagger to extract adjectives and adverbs. These are evaluated using the SO-PMI-IR method to determine the polarity of the comment on various parameters for both faculty and course. The first step involved in the method is the calculation of the Semantic Orientation (SO) of each adjective based on a Hits algorithm. The polarity of the comment is determined as either positive or negative based on a threshold value. This threshold value is calculated by taking the average of the SO values of the positive and negative terms. The final step is the visualization of results.

2.1. Cleaning and Tagging

With the huge volumes of data available on the Internet, the only possible way to extract relevant and useful information is to subject the raw, unstructured text to some kind of analysis which gives an ordered, structured text as output. Data cleaning is the most important of all the preprocessing techniques available. Cleaning the data involves a number of steps. It removes irrelevant data, redundant data, completes the missing values with default values if any and the like. In addition to this the removal of unnecessary punctuation marks, smiley as well. Only the cleaned data can be taken to further analysis. Any data mining or text mining technique applied to the data requires the data to be cleaned. There are a number of methodologies involved for cleaning the data. One of the most popular challenges faced during cleaning is the presence of sms lingos like gud, gr8, osm etc in the comments. In order to be able to help classify these comments into positive or negative, it has to either be converted into its grammatically right forms as good, great and awesome respectively [6]. Once the data is cleaned, the techniques that need to be applied can be done so without having system hang on to some minor fault, like a missing value. This step helps the mining process run smoothly and is hence an extremely vital part of every knowledge extraction process.

Any opinion mining approach involves a tagging section. The tagging is a popular Natural Language Processing technique and is done on a sentence or comment level. Tagging means to take each comment or review provided by users or customers and classify them into their corresponding parts of speech. This is the most commonly used tagger by name Part-Of-Speech (POS) Tagger. Commonly used POS categories are verb, noun, pronoun, adjective, adverb, conjunction, interjection and preposition. The Penn Treebank is a corpus that is said to contain around four point five billion words in English. The POS tagger uses this bank to match with the words in the comments to classify them into different parts of speech [7]. The main objective of using a tagger in any opinion mining process is to extract adjectives or adverbs. Research proves that adjectives and adverbs are good indicators of subjectivity and opinion respectively [10]

2.2. Sentiment Analysis

The one thing that is abundant in this digital era is data. Most of the data available on the Internet can be considered as opinions of people in one way or another. For example, if there is a blog about human rights, the blog consists of data that is solely the opinion of the writer of the blog. Similarly, if there is a review about the latest movie that has hit the box office, it also portrays only the viewpoint of the developer of the particular site. On the other hand if there is a statement on the Internet that says the sun rises in the east it does not mean that it is an individual’s viewpoint. On the contrary, that statement represents a universal truth or a fact. Hence all data available can be classified broadly into two categories, as Facts and Opinions. [1] From the small example provided earlier it can be inferred that extraction of facts is easier than extraction of sentiment or opinions. This is because; a fact will remain a fact wherever it is posted. The statement the sun rises in the east will not differ from one webpage to another. Hence extensive research has been done in the field of extracting facts, whereas extracting opinions is still a relatively new field of research.

Sentiment analysis or opinion mining is a growing field of research that involves attempting to gauge the reaction of the customers and users over a particular product or service and the like. This analysis when done by a company or an organization will have them way ahead in the chart leaving behind their competitors. That is why semantic analysis has become an intense area of research these days. There are three levels in which sentiment analysis is usually done. They are as follows. [2] [3]

- Document level analysis: This level of sentiment analysis assumes that each document conveys opinions about a single entity. It analyzes the document as a whole and determines whether the polarity is positive or negative. Hence this level of analysis is inappropriate for documents which compare more than one entity.
- Sentence level analysis: Each sentence is examined and classified as positive, negative or neutral (no opinion). It points out the difference between two types of sentences namely, subjective and objective sentences. Subjective sentences provide opinions and viewpoints whereas Objective sentences highlight the actual facts expressed in the sentences.
- Entity or Aspect level analysis: This level focuses on the fact that each entity can have one or more aspects involved. And there could be positive opinions about one while there could be a negative opinion about another. For example, a smart phone has a number of aspects like quality of the touch screen, battery life, the camera features and the like. Thus, this level is usually termed a finer-grained analysis.

Semantic analysis has applications not only in the corporate and business world, but also in academia. The main use of semantic analysis is while collecting feedback from the students on both the courses provided and the faculty who teach the courses. Feedback when collected properly with a broad and an open mind shall attempt to answer questions that will help the management improves the standard of education provided by the college or university. Feedback can be collected either in the form of free text or it can be created as a scaling system. The significance of using the former is that it the feedback will prove to be more honest than the numeral ratings and in some cases will help reason the high or low ratings received. [3] The challenging task at hand is to interpret the comments in the same context that they are intended since comments might include abbreviations, smiley and the like to express emotions and opinions.

2.3. Visualization

Visualization of results involves depicting the overall results using different types of graphs and charts like pie charts, bar graphs, line charts, etc. The data present in the work sheets (raw data) is difficult to represent visually. Taking the problem of collecting feedback into consideration, it can be seen that the results of analysis of feedback, that is the comments that are classified as positive and negative cannot be understood fully, be it a corporation or any academic environment unless the whole sheet is read fully. Thus here visualization of results becomes mandatory. R language can be used for data analysis and for creating different kinds of graphs, including bar graphs, scatter plots and word clouds. [8]

3. Implementation

3.1. Proposed System

The proposed system emphasizes on the analysis of the specific comments provided by students on the faculty and the course. The current system takes the feedback in two forms. One is using a rating scheme that has a scale of one to five for rating the various parameters of both the courses and the faculty. The other form is using a comment box that is made optional. The point to be noted here is that, if three students give their ratings as poor, good and excellent the scaling system averages the feedback and gives the overall feedback as good. These results lack in accuracy. The proposed system takes input from the students only in the form of free text. This input is then subjected to a series of steps in analysis and the final results are then depicted in the form of analytical graphs. The results are made available for the faculty and the HODs for improving the standard of education provided.

3.1.1. Advantages

- The results are visualized for better understanding and impact.
- Accuracy of the results is not compromised.

The parameters considered for Faculty are Communication skills, Knowledge Base, Motivating Factor and Practical implementation of concepts. Similarly, the parameters for the courses are Relevance of course, Availability of reading material, Course coverage and Scope for learning new concepts. The students are required to provide their opinions on all these parameters. These comments are then analyzed to obtain the orientation of feedback.

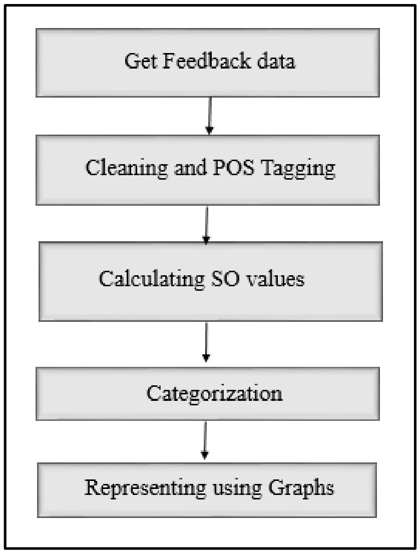


Figure 1

3.2. Data Preprocessing and Tagging

Data preprocessing is the first step involved in any knowledge extraction process. This step focuses on making the large volumes of unstructured data into data that is less unstructured and more convenient for analysis. The steps that are involved generally are removing redundant data, filling in missing values, clearing the irrelevant data and the like. The missing values are usually filled using default values whereas the redundant and irrelevant data are simply eliminated. The feedback that is available at our disposal would be much unstructured and would be unfit for performing analysis as such. Thus the first step towards doing analysis would be to clean the data to get rid of unwanted punctuation marks, smiley and the like. This is generally termed as data preprocessing. The comments are collected from the students and the raw text is given as input to the first module, which is the preprocessing module. This forms the first module of our project. The comments are preprocessed to remove smiley. In addition to that, the popular slang words that are generally used by students like gud, osm, gr8 etc. are also removed in the preprocessing phase. The preprocessed data is then given as input to the next module for further analysis.

As mentioned before there are three levels of sentiment analysis. They are Document level, Sentence level and Aspect level. We have adopted sentence level analysis in which each sentence is taken and analyzed to determine the orientation or polarity of the comments. In order to perform sentence level analysis, the first step done is Part-Of-Speech (POS) tagging. This tagging mechanism involves taking each sentence in the comment and classifying them into various parts of speech. Considering the sentences generally used by people, one can observe that the positivity or negativity of the comments is generally conveyed through the adjectives and adverbs of the sentence. The adverbs and adjectives are hence extracted from the comments and the other parts of speech are ignored. These are then fed into the next module to calculate the Semantic Orientation (SO) of the comments.

3.3. Calculating Semantic Orientation

The basic problem statement in opinion mining can be given as: Given a document or statement A, the sentiment classifier classifies it into either positive or negative category. There have been a number of approaches to achieve such a task: SVM, Naïve-Bayes, kNN and other similar machine learning approaches or by calculating the semantic orientation of the statement by averaging the semantic orientation values of selected words such as adjectives and adverbs. Only these categories of words are used because they are used to express the opinion about any topic or entity (noun). In our project we confined ourselves strictly to sentence level classification of opinions. In our model, we use the SO-PMI-IR method. This method calculates the Semantic Orientation of the given word or phrase and returns a value which is used to judge whether that word or phrase is positive or negative.

The Pointwise Mutual Information (PMI) between two words w1 and w2 is defined as:

$$PMI(w1, w2) = \log_2\left(\frac{P(w1\ \&\ w2)}{P(w1)P(w2)}\right) \qquad (1)$$

Here, P (w1 & w2) is the probability that words w1 and w2 occur together or simultaneously. If the words are statistically independent, then the probability that they co-occur is given by the product P (w1) P (w2). The log of the ratio is the amount of information that we acquire about the presence of one word when we observe the other.

The semantic orientation of a word, *word*, is calculated by SO-PMI-IR as follows:

SO-PMI-IR(word) = PMI(word , {positive words}) – PMI(word, {negative words}) (2)

In the above equation the set of positive words that we have taken are {excellent, good, amazing, positive, right}. Similarly the set of negative words taken are {poor, bad, awful, negative, wrong}.PMI – IR calculates PMI by querying a search engine and recording the number of hits (matching and relevant pages). We used the Google Search Engine for calculating the hits. The Around operator was used while searching to get results within the length of at most 10 words. From the equations (1) and (2), we can obtain the following equation:

$$SO - PMI - IR(word) = \log_2\left(\frac{hits(word\ AROUND(10)positive\_words)\ hits(positive\_words)}{hits(word\ AROUND(10)negative\_words)\ hits(negative\_words)}\right)$$

After the SO of each extracted word is calculated, the average SO for the feedback statement is calculated as a simple average of all the SOs. This average SO of the feedback is compared against a threshold and based on this; the feedback is classified into positive or negative. The threshold value is calculated by taking the average of the SO values of the positive terms and the negative terms. Also, if the word ‘not’ is present in the feedback, then the opposite orientation than the calculated one is assigned to the feedback. Thus we classify all the feedback into ‘positive’ or ‘negative’ category. This method results in nearly accurate classification.

3.4. Categorization

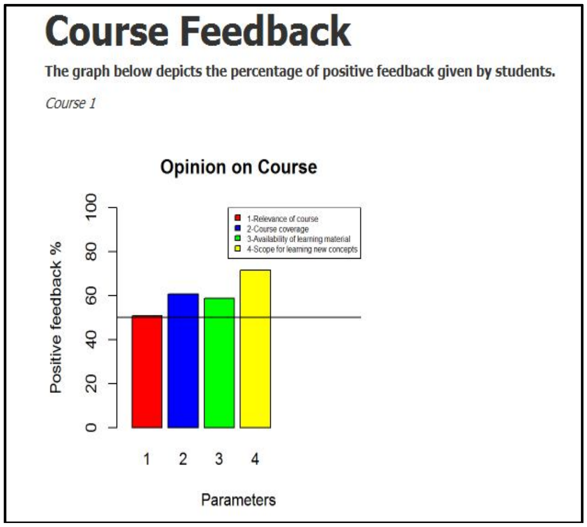
The SO values calculated in the previous module are then used to categorize the comments as positive or negative as follows. A set of positive and negative words are taken as already mentioned. These words are the base with which number of hits of positivity and negativity is calculated. In order to categorize the comments, a threshold value is calculated. This threshold value is determined by taking the SO values of the set of positive and negative words and calculating the average of those values. Thus for each faculty, and for each parameter associated with the faculty the calculated SO values for each adjective or adverb extracted from the comments provided are compared with the threshold value thus determined. The process is similar for all the parameters for the courses as well. If the SO value corresponding to a particular comment is greater than the threshold, it is categorized as positive. Else it is categorized as negative. These results of the analysis phase are then fed into the visualization phase.

3.5. Visualization

Visualization of results is done in the following ways:

- Bar Graphs: These are used to depict the overall percentage of positive feedback received for the faculty and courses along different parameters like Communication Skills, Relevance of Course etc. A threshold line at 50% mark is also included.
- Word Clouds: They represent the most commonly used sentiments (adjectives) to describe the communication skills and the motivating factor of the faculty. The sentiments are extracted from the feedback and represented as word clouds.
- Scatter Plots: These are used to depict the semantic values obtained. They are used to show the strength of the orientation of the feedback. Deliberate negative or positive comments can be identified at outliers, which help in understanding how most of the students gave feedback.

4. Graphs and Charts



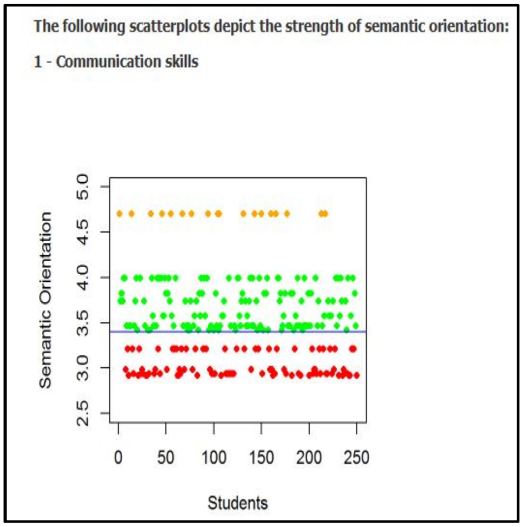


Figure 2: Overall Course Feedback for Course

Figure 3: Scatter Plot between Students and the SO values on the parameter Communication Skills of the Faculty

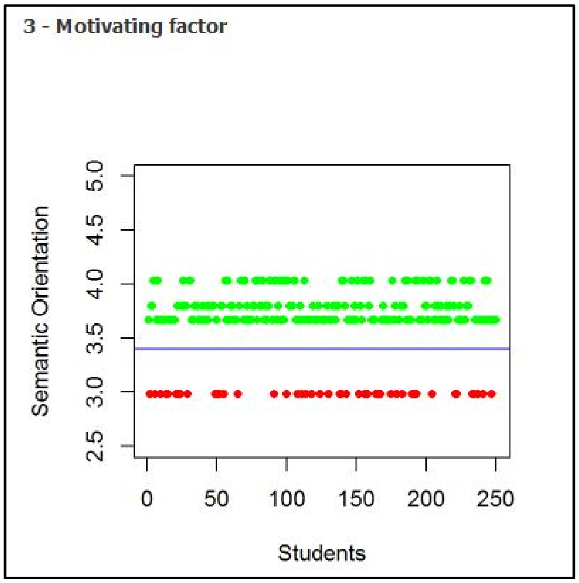


Figure 4: Scatter Plot between values of number of students and SO value of Motivating factor of faculty

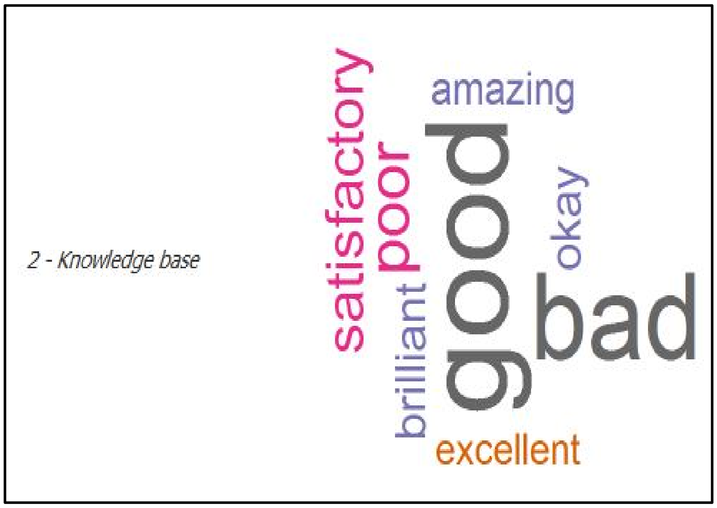
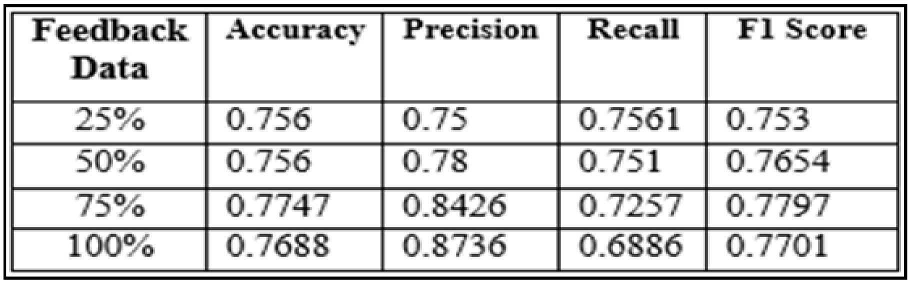


Figure 5: Word Cloud highlighting the frequency of sentiments

5. Experimental Work

The Accuracy, Precision and Recall are calculated for the data. The amount of data taken is varied from 25% to 100% to see how the Accuracy and Precision change for the change in the size of dataset.



Feedback Data	Accuracy	Precision	Recall	F1 Score
25%	0.756	0.75	0.7561	0.753
50%	0.756	0.78	0.751	0.7654
75%	0.7747	0.8426	0.7257	0.7797
100%	0.7688	0.8736	0.6886	0.7701

Table 1

6. Conclusion

We have designed a feedback system which performs opinion mining on the feedback received and displays the results through easy to understand graphs. The feedback is collected using web based form. The results obtained through this method of sentiment mining are mostly accurate. This system is flexible in terms of how the student wishes to give the feedback. We choose to collect feedback in free text to make it more comfortable for students to state their opinions clearly. We also choose to represent the sentiments expressed by the students through word clouds as this enables the faculty to clearly see all the opinions expressed on a particular parameter.

7. References

i. N. Indurkhya & F. J. Damerau (Eds.). (2010) Liu, B, Sentiment analysis and subjectivity, In A Handbook of natural language processing.

ii. Bing Liu (2007). Web Data Mining Exploring Hyperlinks, Contents and Usage Data: Springer-Verlag Berlin Heidelberg.

iii. Bing Liu (2012). Sentiment Analysis and Opinion Mining: Morgan and Claypool Publishers.

iv. Balakrishnan Ramadoss, Rajkumar Kannan (2012). Extracting Features and Sentiment Words from Feedbacks of Learners in Academic Environments, International Conference on Industrial and Intelligent Information (ICIII 2012), IACSIT Press, Singapore.

v. V. K. Singh, P. Kumari, A. Singh, J. Thapa (2011). An Automated Course Feedback System using Opinion Mining, IEEE.

vi. Anjaria, Ram Mohana Reddy Guddeti (2014). Influence Factor Based Opinion Mining of Twitter Data using Supervised Learning, IEEE.

vii. Mitchell Marcus (1993). Building a Large Annotated Corpus of English: The Penn Treebank.

viii. Sameer Bamnote (2014). Data Visualization using R. Cytel Statistical Software and Services.

ix. Turney, Littman (2002). Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. National Council of Research. Canada.

x. Rashid, Asif, Butt, Ashraf (2013). Feature level Opinion mining of Educational Student Feedback Data using Sequential Pattern mining and Association Rule mining. International Journal of Computer Applications, 81, 0975 – 8887