Opinion Mining from Student Feedback

Gokul Ram K, Nanda Kishore V and Shana J
Coimbatore Institute of Technology, Coimbatore, India
gokulramk2019@gmail.com
nandhakishor447@gmail.com
shana@cit.edu.in

Abstract

The field of sentiment analysis is an exciting new research direction thanks to an outsized number of real-world applications where discovering people's opinions is vital in better decision-making. The development of techniques for the documentation-level sentiment analysis is one of the significant components in this area. Recently, people have started expressing their opinions on the online that increased the necessity of analyzing the opinion of online content for various real-world applications. A lot of research can be present for detecting sentiment from the text. But still, there's an enormous scope of improvement of those existing sentiment analysis models. Existing sentiment analysis models can be improved with more semantic and commonsense knowledge.

1.   Introduction

Sentiment analysis is that the process of using natural language processing, text analysis, and statistics to research customer sentiment.At the same time there has been significant research on how sentiment are expressed and plays a vital role in various field like marketing,political,entertainment,feedback . Two methods can be used in sentiment analysis - " Approach based on knowledge " and " Approach based on machine learning". in this paper the machine learning approach is used for the preparation set that builds the classifier to evaluate the sentiments or opinion of the people. Labled data is deployed to the machine learning model and the algorithm gives the desired predefined categories such as whether positive or negative. In text classification labelled data is used as training data to derive the classification model and then it can automatically classify unlabelled text data using the derived model.The goal of the model is used to predict the feedback of the cit college Today we use natural language processing, statistics, and text analysis to extract, and identify the sentiment of words into positive, negative, or neutral categories.The goal of this work is to predict the opinion or sentiment from the feedback texts given by students of the college. This is done by identifying if the sentiment is positive or negative. In this work two different classifier models have been built using machine learning algorithms.To identify the sentiment ,opinion mining can extract the positive or negative.Sentiment analysis helps organisations that unstructured texts come from online sources like email,google form,etc.

2.   Related Work

In [1] author designed a machine learning systems and also developed a data mining concept using artificial intelligence.In [2] author developed an approach in deep learning based on Natural Language Processing, developed a tensor fusion network for multimodal Sentiment Analysis, machine learning approach based on the opinion mining of tweets and machine learning approach for detecting a spam in the twitter tweets.In [3] author analyzed a political sentiment using twitter data and developed a sentimental analysis of twitter data using Naive Bayes algorithm.In [4] author measured a retrieval algorithms of multinomial and multiple - bernoulli models for the sentence retrieval problems and used the both of the models to reduce the syntactical root with the porter stemmer for his experiments.In [5] author developed a Negativity detection in twitter tweets and optimized a decentralized features using machine learning.In [6] author has proposed a Multi-Grained Attention Network (MGAN) model which consists of Input Embedding layer, the Contextual Layer, the Multi Grained Attention Layer and the Output Layer can bring loss in the larger multi-model text and finally he concatenate both the coarse-grained and fine-grained attention vectors as the final representation that is used to determine the sentiment polarityIn [7] author developed the algorithm that is is able to solve the problem space of text-based sentiment analysis that the text given which include special characters or unwanted symbols that can replaced by spaces.In [8] author has obtained a possibility of class whether the given classes results such as success or failure/win or lose,etc… . the algorithms also give the closure accuracy by using this model. In [9] author has developed a capturing favourability using natural language by the concept of sentimental analysis. In [10] author has done natural processing to recognize the sentiment behind the gestures of a specially abled individual.

3. METHODOLOGY

3.1 Dataset Description

The dataset was obtained by collecting feedback about the colleges in Google form as text/short paragraphs. Around 100+ students have given the feedback about college

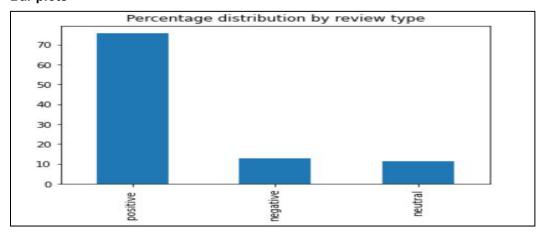### 3.2 Exploratory Data Analysis

Bar plots

Fig.1 Percentage distribution by review type

From Fig 1 it is seen that the percentage of positive, negative and neutral feedback are 80% ,10% ,10% respectively.

Word cloud

Word clouds are the graphical representations of word frequency that give greater prominence to words that appear more frequently during a source text. The larger the word within the visual the more common word was within the dataset.

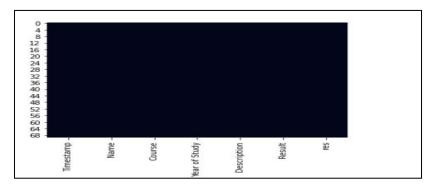**Fig.2 Wordcloud of the Dataset**

Null Values Check

Fig.3 Null values Bar Graph

Inference: This graph shows that there are no null values in the dataset.

3.3 Data preprocessing

Data Cleaning

In this data cleaning process here used two methods Regular expression and 2) Natural language toolkit (nltk).Regular expression is a set of character or pattern is used to find the substring in a set of string.Natural language toolkit is very famous toolkit in text sentiment analysis for the given dataset we have used porterstemmer and stopword .porterstemmer is a method of removing the suffix of the word and bringing it to a base word.stopword are the very common words like 'if', 'but', 'we', 'he','she', and 'they' for removing these words in dataset we use stopword library in nltk.
Here this process is used to change all the caps to small letters and remove the punctuation and clean the text with the special characters.

Fig 4 shows the cleaned data

```
#lets see updated dataset....
corpus

['awesom',
 '',
 'not bad',
 'good',
 'feel great class go realli well compar last sem onlin class',
 'good',
 'good',
 'felt good issu colleg wifi password not given us jio wifi not work work fine st year everi saturday holiday good move us rest
focu non academ activ',
 'grow experi',
 'everyth fine',
 'everyth great',
 'everyth feel comfort',
 'noth',
 'sorri experienc colleg yet',
 'noth',
 'good',
 'everyth go fine onlin class class offlin would much better',
 'mostli onlin interact much',
 'dint experi anyth yet',
 'nice environ',
 'nice',
 'good experi',
 'averag good',
```

Fig : 4 Cleaned Data

Missing Values

In this dataset there are no missing values. It will make the prediction much easier.

Dropped Column

In the csv file timestamp,course,person name, year of study are not required for the prediction it has been dropped

4. MODEL BUILDING PHASES

In order to predict the result and test,the dataset must be split into training and testing subset.The training subset is 70% which means the rest 30% is for testing the model.

4.1. Algorithms Used

Logistic Regression

After creating a 70%(train) 30% (test) . train-test split of the dataset, applied logistic regression which may be a classification algorithm used to solve binary classification problems.Here the logistic regression classifier uses weighted combination of the input features and pass them through a sigmoid function. Sigmoid function transforms any of the real number inputs to variety number between 0 and 1.

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \qquad \ldots\ldots(1.1)$$

In Equation (1.1), it shows that equation of the straight line

$$\frac{y}{1-y}; \text{ 0 for y=0, and infinity for y=1} \qquad \ldots\ldots(1.2)$$

In Equation (1.2) shows that In Logistic Regression y are often between 0 and 1 only, so for this let's divide the above equation by (1-y) .We need range between -[infinity] to +[infinity],and then add the log then it will become.

$$\log\left[\frac{y}{1-y}\right] - b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_n x_n \qquad \ldots\ldots(1.3)$$

In Equation (1.3) shows the final equation of logistic regression.In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X.

Multinomial Naive Bayes

Multinomial naive bayes is that the a part of naive bayes model .The fundamental idea of Naive Bayes is to hunt out the possibility of classes assigned to texts. that Naive Bayes assumes that the presence of a selected feature during a class is unrelated to the presence of the opposite feature and also assumes each feature is given equal importance. Here,feature vectors tell the frequencies of the certain word that has been generated.Naive Bayes needs less training data,less training time and fewer memory consumption. As long as the computed a database of probabilities for terms appearing within the dataset (positive or negative).

$$P(c|x) = \frac{P(x|c)\ P(c)}{P(x)} \qquad \ldots\ldots\ldots(1.4)$$

In Equation (1.4) shows that Bayes theorem calculates probability $P(c|x)$ where c is that the category of the possible outcomes and x is that the given instance which possesses to be classified, representing some certain features.

Gaussian Naive Bayes

Gaussian Naive Bayes could also be a variant of Naive Bayes that follows Gaussian distribution and supports continuous data. We've explored the thought behind Gaussian Naive Bayes.When working with continuous data, an assumption often taken is that the continual value associated with each class are distributed according to a normal (or Gaussian) distribution.

$$P(x_i|y)\ =\ \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \qquad \ldots\ldots\ldots(1.5)$$

5.   RESULTS AND DISCUSSION

Among three models, the multinomial naive bayes have greater accuracy and performance as shown in Table 1.

Model	Accuracy
Multinomial NB	0.85
Gaussian NB	0.70
Logistic regression	0.57

Table 1: Model performance of feedback data

In Table 1 the multinomial NB is performed better than Gaussian NB, Logistic regression within the terms of accuracy with the percentage of 85%(Multinomial NB),70 %(Gaussian NB),57%(Logistic regression).

6.   Conclusion

From the dataset collected our training model says that the feedback of the CIT college is higher than the negative feedback as shown within the count plot.The model summarizes that there are more number of positive feedback about the CIT college.However the result is positive based on the collected data.Other method may be applied in future work to further improve the performance. This article shows a fundamental problem of sentiment analysis, sentiment polarity categorization. Online sites like amazon, flipkart, snapdeal, etc… uses this data for study of the product reviews to get the users attached with all the time.Experiments for both sentence-level categorization and review-level categorization have been performed.

## 7.	References

1)	Simon Tong,"Journal for Machine Learning Research (2014): Need of Machine learning?".

2)	Poria S, Cambria E, Gelbukh A, Bisio F and Hussain A (2015) Sentiment data flow analysis ,"Natural Language Processing toolkit Analyses".

3)	A.Pak, P.Paroubek, Twitter as a corpus for sentiment analysis in machine learning, European Languages Resources Association, 2010.

4)	D. E. Losada, "Language modelling of sentence retrieval: Multiple-Bernoulli and Multinomial models," Bernoulli 2005

5)	Shivani Pathak et al, "Twitter Sentiment analysis using different Algorithms"International Journal for Research in engineering & Engineering Technology (IJRASET)

6)	Fan F, Feng Y, Zhao D (2018) Multi-grained attention network for aspect-level sentiment classification. In: Proceedings of the 2018 conference in natural language processing.

7)	Elin Hanjani Pramitha. (2020). Sentiment Analysis Komentar Pelanggan Hotel Di Purwokerto Menggunakan Naive Bayes Classifier.

8)	Jiang D.,Tao Q.,Wang Z.,Dong L.(2019) Logistic Regression for Sentimental Analysis. In: Patnaik S.,Jain V. (eds)Developments in Intelligent of three Computing.

9)	Nasukawa T. and Yi, J. (2003) "Sentiment analysis: Capturing favorability using natural language processing", K-CAP 2003-Proceedings of the 2nd International Conference on Knowledge Capture, October 23-25, 2003.

10)	Walaa, M., Hassan, A. and Korashy, H. (2014) "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, Vol. 5, No. 4., pp. 1093-1113.

11)	https://monkeylearn.com/blog/sentiment-analysis-machine-learning

12)	https://medium.datadriveninvestor.com/implementing-naive-bayes-for-sentiment-analysis-in-python-951fa8dcd928?gi=9425c7da8f1c