Sentiment Analysis Using Machine Learning Techniques

Rahul Mahajan[1], Sanjeev Jadhav[2], Kedar Deshpande[3], Atharv Jakate[4]
[1,2,3,4]*Department of Information Technology, Government College of Engineering Amravati, Maharashtra, India*

Abstract **- Sentiment analysis is one of the most widely known Natural Language Processing (NLP) task which is also known as Opinion mining or emotion AI. It refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Also it is very well known that people have strong feelings may express them in emotionally laden words. Therefore, programmers use this information to figure out two parameters: First is Sentiment polarity which shows were exactly the feelings lie on positive side or on the negative one. And secondly Sentiment magnitude which depict how strongly the polarity of that sentiment is. To determine these two main factors the algorithm uses either the Dictionary approach or the Categorization approach. Dictionary approach looks up for polarity and magnitude given the word or phrase and reverse it if there is any negation in it. Whereas Categorization approach learn from examples how to categorize any new piece of text using Machine Learning. Modern day sentiment analysis solutions can provide deeper insight. They can capture what specifically people do not like about certain things, and afterward can take steps to fix the issue, or improving a process, they can track how that has improved satisfaction percentage rate. They can also differentiate between feedback that is frequent and feedback that influences satisfaction scores. In this paper we will study about different avenues for sentiment analysis mainly Logistic Regression and Naive Bayes using Hotel Reviews Dataset.**

Index Terms **- Sentiment, Subjectivity, Polarity, Naive Bayes, Logistic Regression, Classification.**

I.INTRODUCTION

Sentiment analysis technique is nothing but a new and efficient kind of text analysis which aims at determining the opinion and subjectivity of reviewers. We can employ it in the field of hospitality like hotels, hospitals etc. but here in this paper we are focusing on Hotel review dataset to be specific. There are hundreds and thousands of hotels we see in your day to day lives. And one of the main problems to tackle is how can one improve its services for which customer sentiment analysis comes into play. To maintain a physical record of all customers who visit the hotel can be a very tedious task and analysis their sentiments is next to impossible. Traditional approach is no good while analyzing these records and so certain machine learning techniques should be employed in order to make the task more efficient.

So, motivation for sentimental analysis is often two-fold. Both consumers and producers highly value customer's opinion about its services. Thus, Sentiment Analysis has seen a considerable effort from industry as well as academia and so we can implement an effective system which will examine customer's sentiments by using the best fit algorithm from the hotels individual dataset.

The Logistic Regression technique (LR) in sentiment analysis first process some text from the given review, which we can represent it as a vector of dimension V, where V corresponds to our vocabulary size. For example, say you have a review, "Hotel services are nice!", then you would put '1' in the corresponding index for any word in the review, and '0' otherwise. As we can see, as V gets larger, the vector becomes more sparse. Furthermore, we end up having many more features and end up training θ V parameters. This could result in larger training time, and large prediction time. Hence, we will extract frequencies of every word and making frequency dictionary. The idea here is to the divide the training set into positive and negative reviews. Count all the words and make a python dictionary of their frequencies in positive and negative reviews.

For every review make a vector of bias unit, sum of all the positive frequencies (words from positive reviews) of all the words and also their negative frequencies.

On the other hand Naive Bayes algorithm (NB) can also be used which is based on Bayes rule, which can be represented as follows:
P(X|Y) =P(Y)P(Y|X) P(X)
In order to predict the sentiment of a review we simply have to sum up the log likelihood of the words in the review along with the log prior. If the value is positive, then the review shows positive sentiment but if the value is negative then the review shows negative sentiment.

To find log prior simply take the log of ratio of number of positive and negative sentiment reviews. i.e.logprior:-
$\log(P(Dpos))-\log(P(Dneg))=\log(Dpos)-\log(Dne\ g)$
Whereas the loglikelihood can be represented as: loglikelihood==$\log(P(Wpos)/P(Wneg))$.
Here $(Wpos)= (freqpos+1)/(Npos+V)$ & $\text{\textdiamond}(Wneg)= (freqneg+1)/(Nneg+V)$
Also $freqpo$S and $freqne$g are the frequencies of that specific word in the positive or negative class. In other words, the positive frequency of a word is the number of times the word is counted with the label of 1. Npos and Nneg are the total number of positive and negative words for all documents (for all reviews), respectively.

II. METHODOLOGY

Algorithms used Logistic regression
The regression is used for predicting the value which is derived from the data sets where we find the independent variables and dependent variables, or we might get series of dependent variables which are eventually get used for regression analysis which decides what could be the future value. So, there are two sorts of regression analysis methods one is linear regression and other one is logistic regression which we intend to use for our sentiment analysis part. Subsequently the reason behind using logistic regression over linear regression is our proposed system have mostly categorical data (Discrete and nominal) through which we can decide whether the customer is happy or not happy with the service. And the precision of log odds can be superior to the linear regression analysis if we provide massive sample of data to train the logistic regression model.

Naïve bayes classifier algorithm:
Naïve bayes algorithm is been proposed to be used as the classification algorithm which does the probabilistic analysis of the information currently available from the previously experienced phenomena or the other itineraries. The naïve bayes classifier uses the bayes theorem and because of that it can make independent assumptions between the concerned features. In our proposed idea of research the prior probability of feature is calculated and subsequently it is used for sentiment analysis.

III. PROPOSED WORK

In the proposed system, we use classification algorithms for sentiment analysis of hotel reviews. This can be done using machine learning. As machine learning techniques easily identify patterns without any human intervention and has continuous improvement ability, they are convenient to be used. Also machine learning has ability to handle large variety of data. For machine learning techniques to be used effectively the data must be noise free.
"Noise" pertains to the stability of the data. Some data is very stable and possesses little variability, while other data swings wildly and unpredictably from one value to another. The degree of that swing is the amount of noise. Noisy data is nothing but meaningless data that cannot be correctly interpreted. It is generated due to faulty data collection, data entry errors etc.

When preprocessing, you have to perform the following:
1. Eliminate handles and URLs
2. Tokenize the string into words.
3. Remove stop words like “and, is, a, on, etc.”
4. Stemming- or convert every word to its stem. Like a dancer, dancing, danced, becomes ‘danc’. You can use porter stemmer to take care of this.
5. Convert all your words to lower case.

A. Machine learning

It is the field of computer science where computers are programmed in such a way that give them capability to learn without being explicitly programmed.
A machine is learning from experiences corresponding to some class of tasks, if its performance in a given task improves with experience.

The process starts with feeding noise free data and then training our machine by building machine learning models using the data and different algorithms. a. The most commonly used Machine Learning algorithms are random forest, support vector machine, k-means clustering, neural network, decision tree.

B. Classification

Classification is. the process of finding a model that describes and distinguishes data classes and concepts. It is the problem of identifying to which of a set of categories (subpopulations), on the basis of a training set of data containing observations, a new observation belongs to. It comes under supervised learning, which learns a function and maps an input to the output. It is an example of pattern recognition. As we have discussed assigning a rating to the hotel based on positive and negative reviews. some of the classification algorithms are logistic regression, Naïve Bayes etc.

Naïve Bayes:

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability and the probabilities of observing various data.

P(h/D) = (P(D/h) * P(h)……(1)

A concept learning algorithm considers a finite hypothesis space 'h' defined over an instance space X.

What can we do if our data d has several attributes?

- Naïve Bayes assumption: - Attributes that describe data instances are conditionally independent given the classification hypothesis.

P(d|h)=P(a1,a2,…….,an|h)= ∑t P(at|h) ... (2)

It is a simplifying assumption, obviously it may be violated in reality

In spite of that, it works well in practice

The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier

One of the most practical learning method Logistic regression:

Logistic Regression is a supervised learning algorithm that is used when the target variable is categorical. Hypothetical function h(x) of linear regression predicts unbounded value. classification problem, where we need to classify whether a hotel review is positive or negative. So, the hypothetical function of linear regression could not be used here to predict as it predicts unbound values, but we have to predict either positive or negative.

1. stigmoid activation function on the hypothetical function of logistic regressions:

h (x) = sigmoid (wx + b) Here, w is the weight vector.
x is the feature vector.
b is the bias.
sigmoid (z) = 1 / (1 + e (- z))

2.the simplified cost function we use:

J = - ylog(h(x)) - (1 - y) log (1 - h(x)) here, y is the real target value
h (x) = sigmoid (wx + b) For y = 0,
J = - log (1 - h(x)) and y = 1,
J = - log(h(x))

3.This cost function is because when we train, we need to maximize the probability by minimizing the loss function.

4.Gradient Descent Calculation: repeat until convergence {
tmpi = wi - alpha * dwi wi = tmpi
}
where alpha is the learning rate.

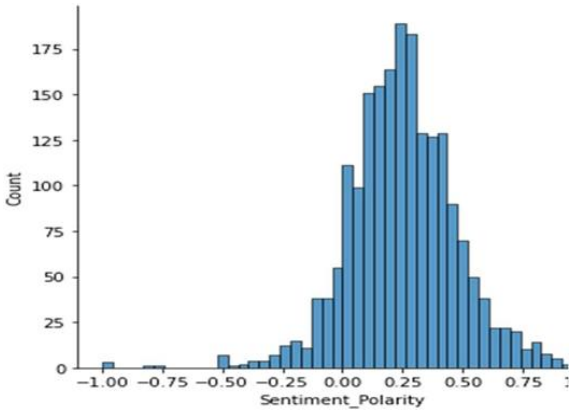
5.The chain rule is used to calculate the gradients like i.e dw.

$$\frac{\partial J}{\partial w} = \frac{\partial J}{\partial a} * \frac{\partial a}{\partial z} * \frac{\partial z}{\partial w}$$

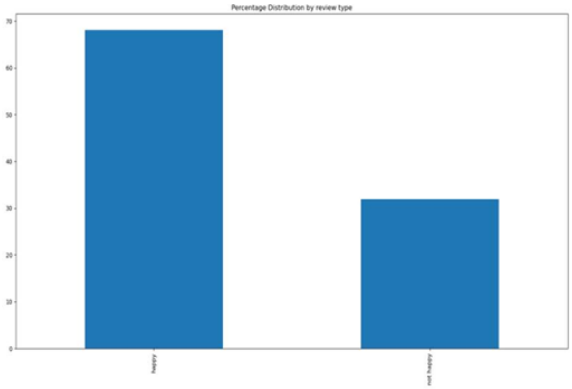
Chain rule for dw
here, a = sigmoid (z) and z = wx + b.

IV.RESULT

Sentiment polarity of data through TextBlob(Using Naïve bayes classfier)



Logistic regression-based sentiment analysis



Description of accuracy, Precision and recall of the naïve bayes classification:

```
# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print ("Confusion Matrix:\n",cm)

# Accuracy, Precision and Recall
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
score1 = accuracy_score(y_test,y_pred)
score2 = precision_score(y_test,y_pred)
score3= recall_score(y_test,y_pred)
print("\n")
print("Accuracy is ",round(score1*100,2),"%")
print("Precision is ",round(score2,2))
print("Recall is ",round(score3,2))
```

```
Confusion Matrix:
 [[119  33]
 [ 34 114]]


Accuracy is  77.67 %
Precision is  0.78
Recall is  0.77
```

Confusion matrix for depicting the accuracy of the classification by logistic regression

```
In [19]: from sklearn.feature_extraction.text import TfidfVectorizer
         from sklearn.linear_model import LogisticRegression

         tvec=TfidfVectorizer()
         clf2=LogisticRegression(solver='lbfgs')

         from sklearn.pipeline import Pipeline

In [20]: model = Pipeline([('vectorizer',tvec),('classifier',clf2)])

         model.fit(IV_train,DV_train)

         from sklearn.metrics import confusion_matrix
         predictions=model.predict(IV_test)
         confusion_matrix(predictions,DV_test)

Out[20]: array([[2417,  304],
                [ 154, 1019]], dtype=int64)

In [22]: from sklearn.metrics import accuracy_score, precision_score,recall_score

         print("Accuracy :", accuracy_score(predictions,DV_test))
         print("Precision :", precision_score(predictions,DV_test,average='weighted'))
         print("Recall : ",recall_score(predictions,DV_test,average='weighted'))

         Accuracy : 0.8823831535695943
         Precision : 0.8889271415963718
         Recall :  0.8823831535695943
```

V.CONCLUSION

The naïve bayes classier implemented through the textblob and using logistic regression for training the model for sentiment analysis is a very approach because we get approximately 80 % accuracy for the sentiment analysis and the case where logistic regression can't be trained without the ample size of categorical data set and also naïve classifier needs the previously available data which is gathered through the experience or any sort of research there we propose the idea to use the lexicon based sentiment analysis approach where we don't need the much information as an prerequisite for sentiment analysis by referring the opinion words and phrases we can discover the sentiment where the retail shop's reviews are very less. Thus combining both these approaches we can achieve a good quality result of opinion summarization.

V. ACKNOWLEDGEMENT

We would like to thank Prof. S.R. Wankhade Madam for her valuable suggestions and help provides throughout the course of analysis, designing and implementing the whole project.

REFERENCES

[1] Bing liu. 2012. "synthesis lectures on human language technologies"
[2] Ewan klein and steven bird. 2009. "Natural Language Processing with Python "
[3] "Sentiment Analysis using TextBlob": https://towardsdatascience.com/my-absolute-go-to- for-sentiment-analysis-textblob-3ac3a11d52
[4] "Natural Language Toolkit3.0.2 documentation": https://www.nltk.org/
[5] "What is Logistic Regression? A Beginner's Guide":https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/
[6] "Complete guide to naïve bayes": https://www.digitalvidya.com/blog/naive-bayes-classifier/
[7] J. W. Pennebaker and R. J. Booth, and M. E. Francis, Linguistic inquiry and word count (LIWC2007) Austin, TX: LIWC (www. liwc. net), 2007.
[8] B. Pang, and L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in

Information Retrieval, vol. 2, no 1-2, pp. 1-135, 2008.

[9] S. M. Kim and E. Hovy, Determining the sentiment of opinions, Proceedings of the 20th international conference on Computational Linguistics, pp. 1367, 2004

[10] J. Golbeck and M. Rothstein, Linking social networks on the web with FOAF: a semantic web case study, Proceedings of the 23rd national conference on Artificial intelligence, pp. 1138-1143, 2008.

[11] M. S. Granovetter, The strength of weak ties,The American journal of sociology, vol. 78, no. 6, pp. 1360-1380, 1973.

[12] S. Rude, and E. M. Gortner, and J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cognition and Emotion, vol. 18, no. 8, pp. 1121-1133, issn 0269-9931, Psychology Press, part of the Taylor and Francis Group, 2004

[13] S. W. Stirman, and J. W. Pennebaker, Word use in the poetry of suicidal and nonsuicidal poets, Psychosomatic Medicine, vol. 63, no. 8, pp. 517, 2001.

[14] J. W. Pennebaker, and A. Graybeal, Patterns of natural language use: Disclosure, personality, and social integration, Current Directions in Psychological Science, vol. 10, no. 3, pp. 90, 2001

[15] M. L. Newman, J. W. Pennebaker, and D. S. Berry, and J. M. Richards, Lying words: Predicting deception from linguistic styles, Personality and Social Psychology Bulletin, vol. 29, no. 5, pp. 665, 2003.

[16] M. R. Mehl and J. W. Pennebaker, The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations, Journal of Personality and Social Psychology, vol. 84, no. 4, pp. 857–870, 2003.

[17] A. Mulac, J. J. Bradac, and P. Gibbons, Empirical support for the gender-as-culture hypothesis, Human Communication Research, vol. 27, no. 1, pp. 121-152, 2001.

[18] L. D. Stone, and J. W. Pennebaker, Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana, Basic and Applied Social Psychology, vol. 24, no. 3, pp. 173-183, 2002.

[19] S. M. Kim and E. Hovy, Determining the sentiment of opinions, Proceedings of the 20th international conference on Computational Linguistics, pp. 1367, 2004.

[20] B. Pang, and L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, vol. 2, no 1-2, pp. 1-135, 2008.

[21] N. Jindal, and B. Liu, Opinion spam and analysis, Proceedings of the international conference on Web search and web data mining, pp. 219-230, 2008.

[22] S. Milgram, The small world problem, Psychology today, vol. 2, no. 1, pp. 60-67, 1967