International Journal on Emerging Technologies **13**(1): 15-21(2022)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Supervised Bernoulli Text Topic Identification Model using Naïve Bayes

Suresh Kumar Sharma[1], Kanchan Jain[1*] and Gurpreet Singh Bawa[2]
[1]Professor, Department ofStatistics, Panjab University, (Chandigarh), India.
[2]Managing Director, Artificial Intelligence, Accenture Chicago Innovation Hub, Madison St, Chicago, United States.

(Corresponding author Kanchan Jain)*
(Received 25 October 2021, Revised 28 November 2021, Accepted 25 December 2021)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In this paper, the concept of document models is conversed with respect to the Bernoulli document approach, that is on basis of the presence or absence of primary blocks of the documents, namely tokens. The research primarily deals with how an unstructured dataset consisting of text documents is converted to structured content with mathematical and statistical foundation and then topic of conversation is predicted (or estimated) based on Bernoulli assumptions. The application of Naïve Bayes approach is discussed for the model under consideration. Examples and sample code snippets in R and Python to execute the same have been included for Bernoulli document model.

Keywords: Text classification, Naïve Bayes, Topic Modelling, Bernoulli Distribution

I. INTRODUCTION

These days, content sharing sites are increasingly gaining importance as a source of information based on behaviour and attitude. Organizations understand economic value of information stored on such sites [1]. There is lot of information put in by users on Twitter, Facebook, online journals, wiki pages, for giving expression to their views. This information pertains to launch of latest electronic gadgets, political events, cricket matches, soccer league results etc. Active users are real and not anonymous and belong to all age groups. Data is poured in by these users instantaneously. A fair idea is obtained about people's opinion and attitudes towards a brand, company or service. For corporations, it also provides a mean for assessing drivers for future business sales [2]. Social sites are being used by about 75% of the internet users and everyday, there is an increase in this percentage [3].

It is easy to have access to data generated from online sources and these data are a potential treasure of information [4]. Getting this information helps in discovering valuable insights in the fields of services, human resources and customer relationship management and marketing. So many companies are focusing on social media for attaining their organizational goals [5].

[2] carried out a statistical analysis on social media through statistical regression and correlation analysis using social unstructured data for prediction of sales in music industry. [6] conducted a study on influence of social media in the financial markets for stock prices.

Sources of unstructured textual data are documents, emails, online forums, electronic news content, blogs, social media feeds and posts, call center logs, customer feedback etc. Text analytics, important in decision making, retrieve information and extract value from the text-based datasets. Statistical Analysis, Computational Linguistics and Machine (Deep) Learning primarily support the idea of Text Analytics. Unstructured textual data from emails, social media, websites need to be processed to be structurally suitable for analysis [7]. Quantification of text data for structural stability can be done using different approaches namely.

Binary/Bernoulli approach: If there is a set of social media posts, and a word of interest, then a simple way to structure the data is to create a flag for each record depending upon whether word of interest is present or not. The random variable of interest representing the presence/absence of a particular characteristic, follows a Bernoulli Distribution. For example, if word of interest is MacBook, then flag or token 1 is assigned to comment 'MacBook is very costly and hence I can't afford to buy a MacBook'' and flag 0 to ''I am going to watch a soccer match today'.

Frequentist approach: Let the same set of documents and same word of interest - 'Macbook' be considered. Then data can be structured by calculating frequencies for each record based on the word of interest. If X counts the number of occurrences of an event, then X follows Poisson Distribution. Using this approach, the comment 'MacBook is very costly and hence I can't afford to buy a MacBook', gets value 2, 'The new MacBook, has better User Interface' has value 1 and 'I am going to watch a soccer match today' gets value 0.

Multinomial approach: This is similar to the Bernoulli approach except the presence flag in the former gets replaced with the frequentist method which takes into account the number of times the tokens or words of interest occur in the text. In this setup, the feature vectors of the document inherently capture the word

frequencies [8] and not merely their occurrence in that document, as in Bernoulli approach. [9] discussed Naïve Bayes approach for a multinomial document model.

The primary objective of this piece of research is to identify uncategorized documents with either of the given categories or labels based on the **Bernoulli distribution** of the terms present in the labelled documents (also known as the training dataset). The task of classifying a piece of text algorithmically does not often require the algorithm to have a deep understanding of the language. This is primarily due to the fact that any text document, irrespective of language, finds a bag-of-words representation whatever be the setting of the algorithm in the background. Here the bag resembles a collection having elements where repetition is allowed. This representation is very simple and intuitive and simply focuses on the words occurring in the text document and the frequency. As a result, the notion of ordering of the words or the arrangement in which it existed in the original document, gets ignored. Let there be a document *D* having class/topic/category denoted by *C*. Let the various realizations of *C* be denoted as C_1, C_2 etc. The posterior probability $P(C \mid D)$ is given by

$$P(C \mid D) = \frac{P(D \mid C) \times P(C)}{P(D)} \propto P(D \mid C)P(C) \quad (1)$$

Under the assumption of Naïve Bayes [10], (Qin, Tang and Chen, 2012), Bernoulli document model will be discussed in this research. It must be noted that it follows a bag-of-words representation for the documents. The documents are represented in the model with the help of feature vectors, the components of which are the types of words occurring in the documents. Let us assume that there is a vocabulary *V* over the given documents, which contains $|V|$ types of words. This implies that the dimension of the feature vector *D* equals the count of word types $|V|$. In Bernoulli document model, documents find representation using feature vectors with Boolean or binary components assuming realization 1 if the equivalent word occurs in the document and 0 if the word is absent. Similarly, for Multinomial document model, documents find representation using feature vectors with integer components having values denoting the frequencies of the associated words in the concerning documents. In Section II, we describe the Bernoulli model setup in detail and give an example in Section III. In Section IV, we show the application of Bernoulli document model on a real-life data and its implementation in R. Implementation in Python is provided in same section which is followed by conclusions in Section V.

II. BERNOULLI MODEL

A document, in a Bernoulli set up, finds representation using a vector of binary type, which in turn is a representation of a point in the word space. Let there be a vocabulary *V* having $|V|$ words. The t^{th} item of a document vector will correspond to word w_t of the vocabulary.

If *b* represents document *D*'s feature vector; then b_t, the t^{th} item of *b*, assumes the value 0 or 1 depending upon the non-occurrence/occurrence of word w_t within the document text.

As an example, let there be a vocabulary:

V = {blue, green, dog, tiger, biscuit, banana}

Cardinality of *V* or $|V|$= dimension of feature vector *D* = 6.

To illustrate further, consider a document "the blue dog ate a blue biscuit". Let d^B be the associated Bernoulli feature vector, and d^M the multinomial feature vector. Thus

$$d^B = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \qquad d^M = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Hence, for classifying the document, (1) can be used. Now that requires estimation of the document's likelihood conditional on the class/topic/category *C*, $P(D|C)$ and the associated prior probability $P(C)$ of the classes. The Naïve Bayes assumptions are applicable to either of the two document models that would be used, while estimating the likelihood $P(D|C)$.

Under the assumption of Naïve Bayes which suggests that event of any word being present in the document is not dependent on presence of any other word, enables us to express the likelihood $P(D|C)$ of the document in terms of the distinct solitary word likelihoods $P(w_t|C_k)$ as

$$P(D \mid C_k) = P(b \mid C_k) = \prod_{t=1}^{|V|} [b_t \times P(w_t \mid C_k) + (1 - b_t) \times (1 - P(w_t \mid C_k))] \quad (2)$$

where $P(w_t|C_k)$ is the probability that word w_t occurs in a document belonging to k^{th} category or topic and $(1 - P(w_t|C_k))$ is the probability that w_t does not occur in some document belonging to this category (2) iterates over every word present in the vocabulary.

If word w_t occurs, then b_t equals 1 and associated probability is $P(w_t|C_k)$. Otherwise, b_t equals zero with associated probability as $(1 - P(w_t|C_k))$. This can be looked upon as a model to generate feature vectors of documents belonging to class *k*, where the feature vector of the document is exhibited as a collection of coin tosses with $|V|$ weights, where t^{th} toss has success probability as $P(w_t|C_k)$.

Let $n_k(w_t)$ be the count of documents of category *k* where w_t is present and let N_k be the total count of documents belonging to the k^{th} class. Then

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{N_k} \quad (3)$$

represents relative frequency of documents belonging to category *k* and containing w_t.

For a given training set having *N* documents, the prior probability for category *k* is written as

$$\hat{P}(C_k) = \frac{N_k}{N} \quad (4)$$

Therefore, for a given training dataset consisting of labeled documents, each associated to either of the *k* categories, a Bernoulli classification model can be estimated in the following manner:

1. Define the vocabulary *V* where the count of words in it provides the feature vector dimensionality;

— In the training dataset, enumerate

— N, the total count of documents in training dataset
— N_k, the count of documents with category labels k, where k ranges from 1 to K
— $n_k(w_t)$, the count of documents belonging to category k, and having word w_t where k ranges from 1 to K and t ranges from 1 to $|V|$
2. Estimate the likelihood $P(w_t|C_k)$ using (3);
3. Estimate the prior probabilities $P(C_k)$ using (4).
Finally, for classifying an unknown and unseen document D, the posterior probability needs to be estimated for each category k using the combination of (1) with the Bernoulli model document likelihood equation as

$$P(C_k \mid b) \propto P(b \mid C_k) \times P(C_k)$$
$$\propto P(C_k)\left[\prod_{t=1}^{|V|}[b_t \times P(w_t \mid C_k) + (1-b_t)\times(1-P(w_t \mid C_k))]\right] \quad (5)$$

III. EXAMPLE OF BERNOULLI MODEL

Let there be a collection of documents, every one of which belongs to one of the two topics, that is, Sports or Informatics, denoted by S and I respectively. Now, for a given training dataset having eleven documents, the objective is establishing an estimation for a Bernoulli document classifier, to label unseen documents pertaining to Sports or Informatics.

Let the vocabulary V consist of 8 words as

$$V = \begin{bmatrix} w_1 = goal \\ w_2 = tutor \\ w_3 = \text{var}iance \\ w_4 = speed \\ w_5 = drink \\ w_6 = defence \\ w_7 = performance \\ w_8 = field \end{bmatrix}$$

Hence, all the documents can have a representation in form of a vector of 8 dimensions. A document D_i can now be denoted as a row vector m_i where m_{it} denotes the count of word w_t in D_i.
The training dataset is shown below in the form of a matrix corresponding to each category or topic where each row signifies a document vector of 8 dimensions.

$$B^{Sport} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$B^{Inf} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Now, the objective is classifying the following vectors
1. $b_1 = (1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1)^T$
2. $b_2 = (0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0)^T$

into either of the topics with the help of a Naïve Bayes (NB) classifier.
Enumerations in the training dataset are made as $N = 11$, $N_S = 6$, $N_I = 5$.

The prior probabilities can be estimated from the training dataset using (4) and are given by

$$\hat{P}(S) = \frac{6}{11}; \quad \hat{P}(I) = \frac{5}{11}.$$

The count of documents $n_k(w)$ in the training dataset, and estimates of word likelihoods are given in the following table:

Table 1: Count of Documents and Estimates of Word Likelihood

	$n_s(w)$	$\hat{P}(w \mid S)$	$n_1(w)$	$\hat{P}(w \mid I)$
W_1	3	$\frac{3}{6}$	1	$\frac{1}{5}$
W_2	1	$\frac{1}{6}$	3	$\frac{4}{5}$
W_3	2	$\frac{2}{6}$	3	$\frac{3}{5}$
W_4	3	$\frac{3}{6}$	1	$\frac{1}{5}$
W_5	3	$\frac{3}{6}$	1	$\frac{1}{5}$
W_6	4	$\frac{3}{6}$	1	$\frac{2}{5}$
W_7	4	$\frac{4}{6}$	3	$\frac{3}{5}$
W_8	4	$\frac{4}{6}$	1	$\frac{1}{5}$

Posterior probabilities of two test data points are computed for purpose of classification.

$$\hat{P}(S \mid b_1) \propto \hat{P}(S) \times \prod_{t=1}^{8}[b_{1t} \times \hat{P}(w_t \mid S) + (1-b_{1t}) \times (1-\hat{P}(w_t \mid S))]$$
$$\propto \frac{6}{11}\left(\frac{1}{2}\times\frac{5}{6}\times\frac{2}{3}\times\frac{1}{2}\times\frac{1}{2}\times\frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\right) = \frac{5}{891}$$
$$\approx 5.6\times10^{-3}$$
$$\hat{P}(I \mid b_1) \propto \hat{P}(I) \times \prod_{t=1}^{8}[b_{1t} \times \hat{P}(w_t \mid I) + (1-b_{1t}) \times (1-\hat{P}(w_t \mid I))]$$
$$\propto \frac{5}{11}.\left(\frac{1}{5}.\frac{2}{5}.\frac{2}{5}.\frac{1}{5}.\frac{1}{5}.\frac{1}{5}.\frac{2}{5}.\frac{1}{5}\right) = \frac{8}{859375}$$
$$\approx 9.3\times10^{-6}.$$

Hence, b_1 can be categorized as belonging to S.

$$\hat{P}(S \mid b_2) \propto \hat{P}(S) \times \prod_{t=1}^{8}[b_{2t} \times \hat{P}(w_t \mid S) + (1-b_{2t}) \times (1-\hat{P}(w_t \mid S))]$$
$$\propto \frac{6}{11}.\left(\frac{1}{2}.\frac{1}{6}.\frac{1}{3}.\frac{1}{2}.\frac{1}{2}.\frac{1}{3}.\frac{2}{3}.\frac{1}{3}\right) = \frac{12}{42768}$$
$$\approx 2.81\times10^{-4}.$$

$$\hat{P}(I \mid b_2) \propto \hat{P}(I) \times \prod_{t=1}^{8} [b_{2t} \times \hat{P}(w_t \mid I) + (1 - b_{2t}) \times (1 - \hat{P}(w_t \mid I))]$$

$$\propto \frac{5}{11} \cdot \left(\frac{4}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{4}{5} \right) = \frac{34560}{4296875}$$

$$\approx 8.1 \times 10^{-3}.$$

This implies that $b_2$can be categorized as belonging to I.

IV. REAL-LIFE IMPLEMENTATION IN R

In order to implement Bernoulli Naïve Bayes on text data in R, the following line of action can be adopted.

First, the required packages are imported into the R environment.

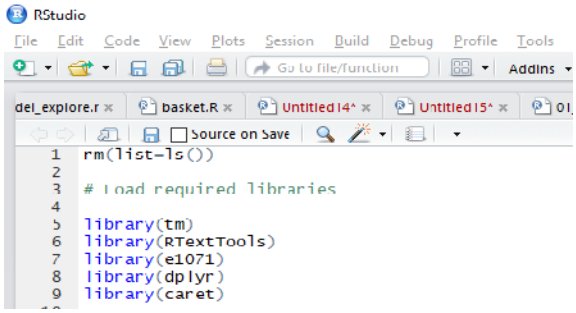


Fig. 1. Code to Import Required Packages.

Next, the dataset is imported into the environment using the code shown in Fig. 2. The used dataset is the Cornell IMDB movie reviews dataset. There are reviews of 2000 movies and an associated positive or negative label based on the sentiment.

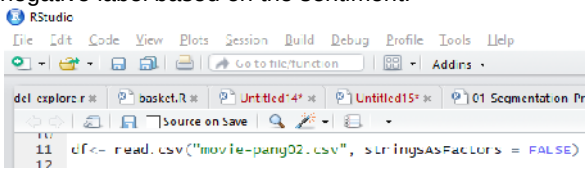


Fig. 2. Importing Dataset into R.

The imported dataset looks as in Fig. 3.

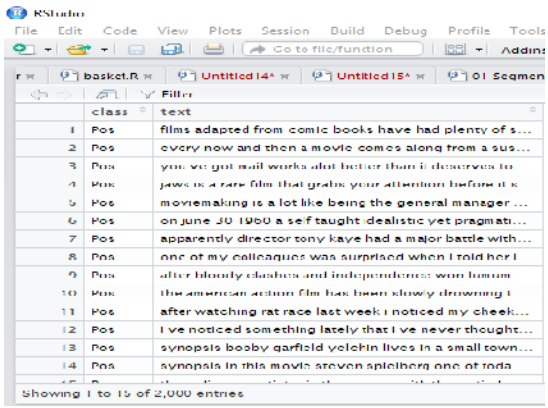


Fig. 3. Snapshot of Dataset.

Then the ordering of the dataset is randomized twice to ensure that any kind of patterns in labels are not present while getting the train and test split.

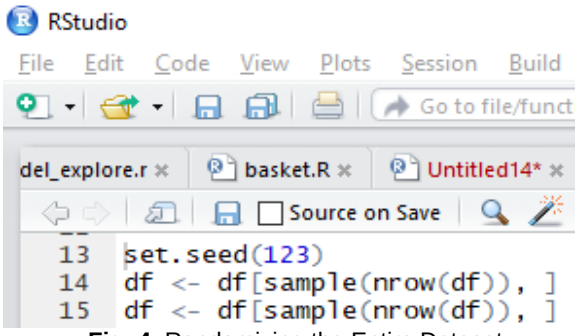


Fig. 4. Randomizing the Entire Dataset.

The dataset now looks as in Fig. 5 and when compared to the earlier data snapshot (Fig. 3), it is different in ordering.

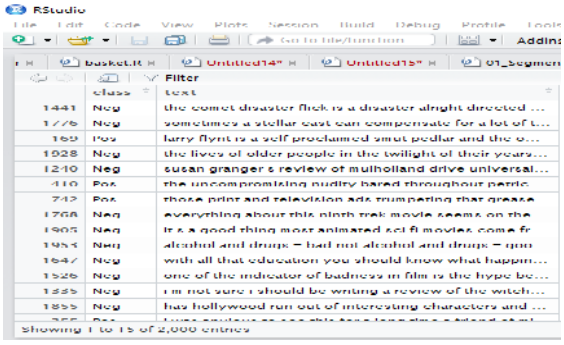


Fig. 5. Randomized Dataset.

The 'class' variable is converted to type 'factor' and corpus of the 'text' variable is created.

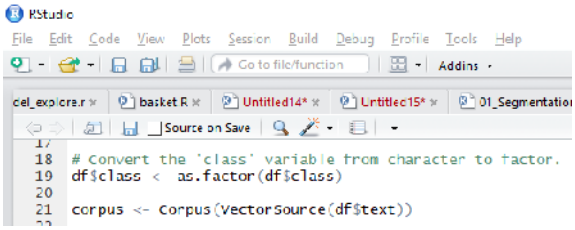


Fig. 6. Code for Conversion of 'class' to Factor and Creation of Corpus of 'text'.

Once the corpus is created, it is cleaned by data pre-processing such as conversion to lower case, removing punctuations, removing numbers, removing stopwords and clearing leading whitespaces.

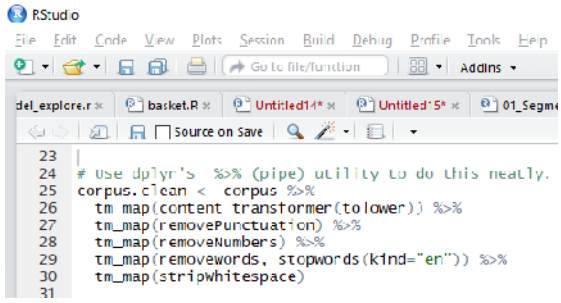


Fig. 7. Preprocessing Step.

Once the pre-processing step is completed, a Document Term Matrix on the clean corpus is created.

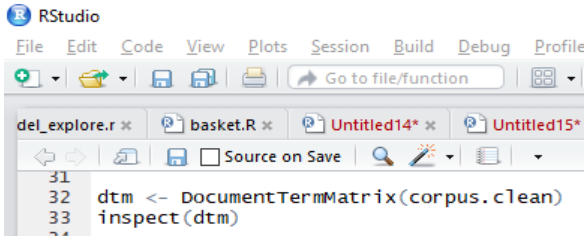


Fig. 8. DTM Creation on Cleaned Corpus.

On inspecting the Document Term Matrix (DTM), the observed information is shown in Fig. 9.

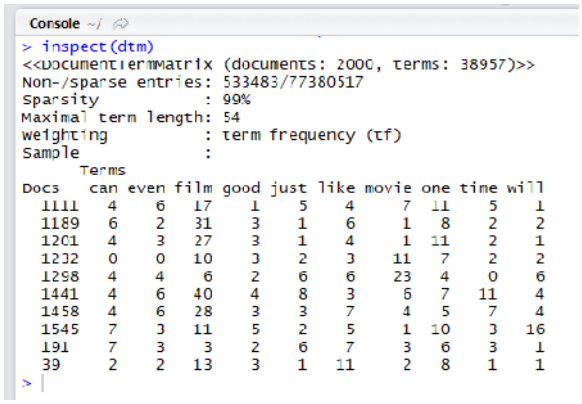


Fig. 9. Resultant DTM Built on Cleaned Corpus.

It is seen that number of terms, which are features, is quite high (38957). In order to reduce them, the terms which occur at least in 5 or more documents would be considered for the analysis. Thus, a dictionary of only those terms is created and the DTM is constructed on that basis.

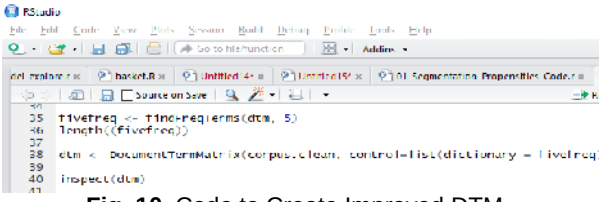


Fig. 10. Code to Create Improved DTM.

Since the dataset was already randomized, the first 1500 rows are taken as train and remaining 500 as test.

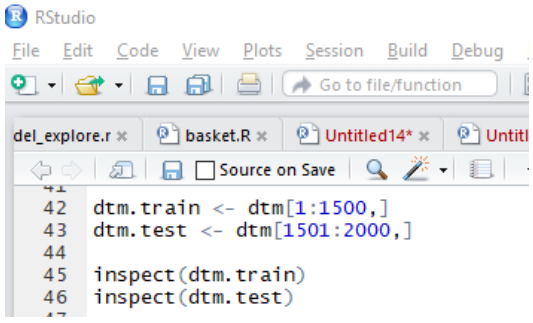


Fig. 11. Train Test Split.

Now the DTM objects are converted back to data frames for ease of handling and this is depicted in Fig. 12.

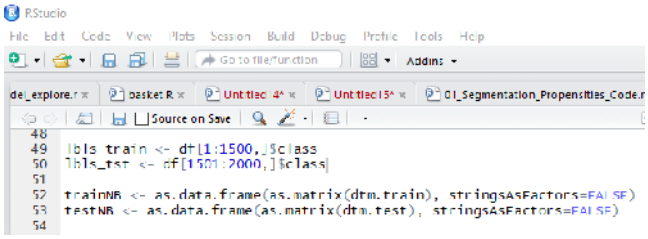


Fig. 12. DTM to Data Frame Conversion.

Next, since the underlying distribution is Bernoulli, the DTM should contain each feature in the form of a 0-1 factor.

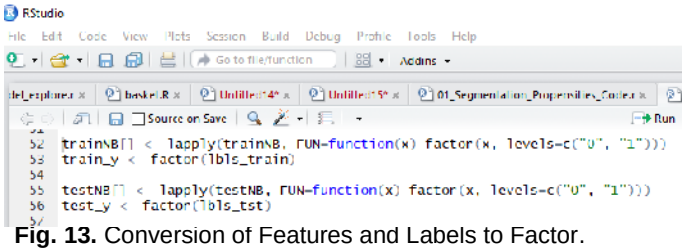


Fig. 13. Conversion of Features and Labels to Factor.

Finally, Bernoulli Document model is fitted on the training dataset and used to make predictions on the test set.

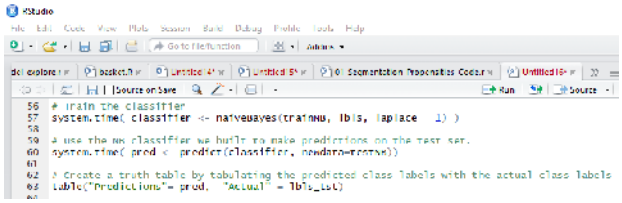
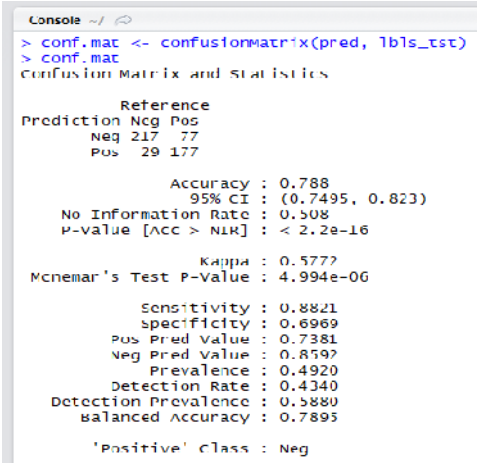


Fig. 14. Fitting Model and Predicting.

The Confusion Matrix function is used to build the confusion matrix and get relevant statistic as shown in Table 2.

Table 2: Confusion Matrix and Relevant Statistics.



In Table 2, the fitted model shows an accuracy of ~ 79% in classifying the unlabeled observations. Other model diagnostics are as below:

1. Accuracy of 79% implies closeness of the sample statistic to the population parameter.

2. 95% Confidence Interval (CI) is (0.7495, 0.823). It is a combination of the estimates of intervals and probabilities. It implies that if the identical sampling approach is utilized for selecting distinct samples and an interval estimate is calculated for each of them, then the actual population parameter can be expected to be within the interval estimates for approximately 95% of times.

3. No Information Rate, the finest approximation conditional that zero information beyond the complete class distribution is provided, is 0.508.

4. Kappa: In the task where two binary variables are attempted by two entities in measuring the identical object, Cohen's Kappa (or simply Kappa) [11] can be used as an agreement measure between them. Its value is always $\leq$ 1. A realization of Kappa 1 suggests agreement in the perfect sense and accordingly for less than 1,

$Poor: < 0.20$

$Fair: 0.20 \leq Kappa \leq 0.40$

$Moderate: 0.40 \leq Kappa \leq 0.60$

$Good: 0.60 \leq Kappa \leq 0.80$

$Very\ good: 0.80 \leq Kappa \leq 1.00$

In our case, Kappa takes the value 0.5772, which lies in the moderate range.

5. McNemar's Test p-Value [12]: A small p-value suggests evidences of association.

For our example, it equals $4.994*10^{-6}$ which shows an association between dependent and independent variable.

6. Sensitivity, the ability of the test to make correct true positive identification is 0.8821 which is very good.

7. Specificity, the ability of the test to make correct true negative identification, is 0.6969.

8. Pred Value: Positive Pred Value (PPV) and Negative Pred Value (NPV) are respectively, the proportions of positive and negative outcomes that are True Positive (TP) and True Negative (TN).

PPV is 0.7381 and NPV is 0.8592.

8. Prevalence: Here prevalence, share of cases in the given population at an instance, is 0.4920.

Above model diagnostics show that the fitted model is classifying the documents quite well.

A simple implementation of the Bernoulli Naïve Bayes model in Python is done using the following code:

```
# Load libraries
import numpy as np
from sklearn.naive_bayes import BernoulliNB

# Create three binary features
X = np.random.randint(2, size=(100, 3))

# Create a binary target vector
y = np.random.randint(2, size=(100, 1)).ravel()

# View first ten observations
X[0:10]

# Create Bernoulli Naive Bayes object with prior probabilities of each class
clf = BernoulliNB(class_prior=[0.25, 0.5])

# Train model
model = clf.fit(X, y)
```

V. CONCLUSION

In this research, it is seen how categorization of unlabeled documents can be done using underlying Bernoulli distribution for words, based on the posterior probabilities obtained from the pre-labeled training dataset. This approach, which is lexical in nature, deals with the features obtained from the labeled training dataset only and focuses on just the presence/absence of words across the documents. Thus, in this way, the supervised approach classifies the unlabeled documents to either of the categories under study.

VI. FUTURE SCOPE

Models will be suggested for audio data by effectively converting the data in audio format to text format through statistical driven algorithms.

Acknowledgement. The authors are grateful to the worthy editor and referee for their valuable comments and suggestions which have led to an improvement in the manuscript.

Conflict of Interests. It is declared that there are no conflicting interests among all three authors.

REFERENCES

[1]. Ghose, A. and Panagiotos, I. (2010). The economizing project at nyu: Studying the economic value of user-generated content on the internet. *Journal of Revenue and Pricing Management, 8*: pp 241–246.

[2]. Dhar, V. and Chang, E. A. (2009). Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive Marketing, 23*(4): pp 300–307.

[3]. Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of socialmedia. *Business Horizons, 53*(1): pp 59–68.

[4] .Dey, L. and Haque, S. M. (2008). Opinion mining from noisy text data. In Proceedings of the second workshop on analytics for noisy unstructured text data, pp 83–90.

[5]. Murdough, C. (2009). Social media measurement: It's not impossible. *Journal of Interactive Advertising, 10*: pp 94–99.

[6]. Tirunillai, S. and Tellis, G. (2012). Does chatter really matter? dynamics of user-generated content and stock performance. *Marketing Science, 31*: pp 198–215.

[7]. Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M. and Marinescu, V. (2013). Converting unstructured and semi-structured data into knowledge. Proceedings - RoEduNet IEEE International Conference, Romania.

[8]. Manning, D. C., Raghavan, P. and Schutze, H.

(2008). Introduction to Information Retrieval. *Cambridge University Press*, pp 253 – 280.
[9]. Jain, K., Sharma S. K. and Bawa, G. S. (2022). Supervised Multinomial Text Topic Identification using Naïve Bayes. *Asian Journal of Statistical Scences* (Accepted for publication).
[10]. Qin, F., Tang, X. and Cheng, Z. (2012). Application and research of multi-label Naïve Bayes Classifier, *Proceedings of the 10th World Congress on Intelligent Control and Automation,* Beijing, pp 764-768.
[11]. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement. 20*(1): 37–46.
[12]. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2): 153–157.

How to cite this article: Suresh Kumar Sharma, Kanchan Jain and Gurpreet Singh Bawa (2022). Supervised Bernoulli Text Topic Identification Model using Naïve Bayes. *International Journal on Emerging Technologies, 13*(1): 15–21.