**Natural Language Processing**

**Subject Code: CSE4022**

**Slot: E2+TE2**

Final Project Review

**NEWS CATEGORIZATION USING MULTINOMIAL NAÏVE BAYES ALGORITHM**



**Prepared by:**

**Samartha Pal (18BCE2434)**

**Hansraj Kumar Rouniyar (18BCE2472)**

**Kamal Subedi (18BCE2479)**

**Ashmit Bhatta (18BCE2486)**

**Submitted to:**

**Respected Dr.Sharmila Banu K Ma’am**

**Table of Contents** Abstract...................................................................................................3

1.Introduction.........................................................................................3,41.1Problem Statement..............................................................................4

2. Literature Survey.............................................................................5

2.1 News Classification Using Naïve Baye’s Classifier...................5

2.2 Text News Classification System using Naïve Bayes Classifier......6

2.3 Sentiment Analysis Using Naïve Bayes Classifier.................6-7

2.4 Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification.......................................................7-8

2.5 Document Classification of Assamese Text Using Naïve Bayes Approach.................................................................8-9

2.6Some Effective Techniques for Naive Bayes Text Classification...9

3. Tabulated Form of survey...............................................................9

4. Methodology...................................................................................12

5. DataSet Source............................................................................ 13

6. Implementation...........................................................................13-22

7. Result..........................................................................................22

8. Refrences...................................................................................22-24

**ABSTRACT**

Sentiment analysis is a kind of text classification that catalogs texts based on the sentiment orientation of opinions they contain. It thus plays an important part of Natural Language Processing. This field is particularly of use to merchants, stock traders, and in election works. Automatic sorting and categorization of context has been playing a vital role lately and is making everyone’s lives easier and sorted, technologically. One of the many Sentiment Analysis applications-text classification, in this case, News Categorization will be employed for given set of news headlines using Multinomial Naïve Bayes classifier, a sub branch of machine learning based sentiment analysis.

**1)Introduction**

[Naïve Bayes](https://www.mygreatlearning.com/blog/introduction-to-naive-bayes/), which is computationally very efficient and easy to implement, is a learning algorithm frequently used in text classification problems. Two event models are commonly used:

* Multivariate Bernoulli Event Model
* Multivariate Event Model

The Multivariate Event model is referred to as Multinomial Naive Bayes.(*which we are using in our project here)*

When most people want to learn about Naive Bayes, they want to learn about the Multinomial Naive Bayes Classifier. However, there is another commonly used version of Naïve Bayes, called Gaussian Naive Bayes Classification.

Naive Bayes is based on Bayes’ theorem, where the adjective Naïve says that features in the dataset are mutually independent. Occurrence of one feature does not affect the probability of occurrence of the other feature. For small sample sizes, Naïve Bayes can outperform the most powerful alternatives. Being relatively robust, easy to implement, fast, and accurate, it is used in many different fields.

For a simple Example, Spam filtering in email, Diagnosis of diseases, making decisions about treatment, Classification of RNA sequences in taxonomic studies, to name a few. However, we have to keep in mind about the type of data and the type of problem to be solved that dictates which classification model we want to choose. Strong violations of the independence assumptions and non-classification problems can lead to poor performance. In practice, it is recommended to use different classification models on the same dataset and then consider the performance, aswell as computational efficiency.

1.1 Problem Statement

In practice, the conditional independence assumption in Naïve Bayes classifier is rarely true, and as a result, its probability estimates are often suboptimal. In order to reduce inaccuracies from Naive assumption, many approaches are proposed in literature. Compared with classical methods, Bayesian methods provide a natural and principled way of combining prior information with data, within a solid decision theoretical framework. One can incorporate past information about a parameter and form a prior distribution for future analysis.

**2. Literature Survey**

Here we have overviewed 5 significant papers so as to give various different investigations and applications of Naïve Bayes related surveys and their significant real-life applications, portraying how individual logical thoughts of creators can get together through cooperation to cause a blast of logical discoveries.

**2.1.**

***News Classification Using Naïve Baye’s Classifier***

**Publisher** : International Journal of Advanced Research in Computer Science and Software Engineering

**Author** : Suraj Patidar

With the RAPID growing rate of techniques for manipulation in Real Time Data. News classification has increased the interest in the research of text mining. Correctly identifying the news into particular category is still presenting challenge because of large and vast amount of features in the dataset. In regards to the existing classifying approaches, Naïve Baye’s is potentially good at serving as a document classification model due to its simplicity. This paper proposed the news classification using Naïve Baye’s classifier in which several types ofdifferent news has been classified like politics, business, entertainment and health. The whole implementation has been taken place in Visual Basic 2010 by using language C#.

**2.2**

***Text News Classification System using Naïve Bayes Classifier***

**Publisher:** An International Journal of Engineering Sciences (Punjab University)

**Authors** : Shruti Bajaj Mangal and Dr Vishal Goyal

This paper describes the Naive Bayes text News classification system developed for Punjabi Language. News corpus is used for training and testing purpose of the classifiers. Language specific preprocessing techniques are applied on raw data to generate a standardized and reduced-feature lexicon. Punjabi language is morphological rich language which makes those tasks complex. Statistical characteristics of corpus and lexicon are measured which show satisfactory results of text pre-processing module. We are able to get satisfactory results using Naive Bayes Classifier.

**2.3**

**S*entiment Analysis Using Naïve Bayes Classifier***

**Publisher:** International Journal of Innovative Technology and Exploring Engineering (IJITEE)

**Author :** Manish Sharma

Twitter is a web service and social communication platform which allow users to address their tweets in different domains. Public can easily and efficiently explicit their perspectives and ideas on a wide variety of cluster on topics via social networking websites. As online data is abundant through different platforms like social networks, twitter, Facebook, etc... Analysing the data is of paramount importance in drawing inference from the data. Hence, in our research, we try to perform sentiment analysis on twitter data by using a Naive Bayesian algorithm. By using our model, we can measure the customers opinions and perceptions and can be enhanced to any desired level depending on the data gathered from on line resource.

**2.4**

***Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification***

**Publisher:** ICACTM

**Author :** Gurinder Singh

Document/Text Classification has become an important area in the field of Machine Learning. On account of its wide applications in business, ham/spam filtering, health, e-commerce, social media sentiment, product sentiment among customers etc., various approaches have been devised to accurately predict the category or to classify any of the new text/document under consideration. Nowadays, news articles in the newspaper present various kinds of sentiments or inclination of the news article towards a negative or positive sentiment and hence, the content of the news can actively be used to judge the impact on the reader. The paper aims to predict that whether the sentiment of the news article is positive or negative using the two popular approaches of Naïve Bayes Text Categorization i.e.

Multivariate Bernoulli Naïve Bayes Classification and Multinomial Naïve Bayes Classification. Also, the research aims to identify that which approach between the given two approaches perform better for the given dataset.

**2.5**

***Document Classification of Assamese Text Using Naïve Bayes Approach***

**Publisher** : International Journal of Computer Trends and Technology (IJCTT)

**Authors** : Moromi Sharma and Shahin Akhter

Document classification has become an emerging technique in the field of research due to the abundance of documents available in digital form. Document classification can be used to organize data into smaller and meaningful classes. Correctly identifying a document into a particular class is still a huge challenge particularly in Assamese text as very few work has been done in this field . In this paper we have done document classification using Naïve bayes classifier. In regards to the various classifying approaches, Naïve Bayes is potentially good at serving as a document classification model due to its simplicity.

The aim of this paper is to highlight the performance of employing Naïve Bayes in document classification. In this paper the document is classified into one of the four classes i.e. sports, politics , law and science. To build and evaluate the classification model, a total 200 documents is split into two datasets, namely training set and testing set, in which 60% of the documents is used as training set whereas the remaining 40% is used as the testing set. The results have been validated using statistical measures of precision , recall and their combination F-measure. Results show that Naïve Bayes is a good classifiers .

**2.6**

***Some Effective Techniques for Naive Bayes Text Classification***

**Publisher :** IEE Transactions on knowledge and data engineering

**Author :** Sang-Bum Kim and Hae Chang Rim

While naive Bayes is quite effective in various data mining tasks, it shows a disappointing result in the automatic text classification problem. Based on the observation of naive Bayes for the natural language text, we found a serious problem in the parameter estimation process, which causes poor results in text classification domain. In this paper, we propose two empirical heuristics: per-document text normalization and feature weighting method. While these are somewhat ad hoc methods, our proposed naive Bayes text classifier performs very well in the standard benchmark collections, competing with state-of-the-art text classifiers based on a highly complex learning method such as SVM.

**3.TABULATED FORM :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Serial | Topic | Methodology | Result | Critical |
| no. |  |  |  | Analysis |
|  | **News** | Naive Baye’s | Used naive | The whole |
| 1 | **Classification** | classifier | Baye’s | implementati |
|  | **Using Naïve** |  | classifier in | on has been |
|  | **Baye’s** |  | which | taken place in |
|  | **Classifier** |  | several | Visual Basic |
|  | Publisher: |  | types of | 2010 by |
|  | International |  | different | using |
|  | Journal of |  | news. | language C# |
|  | Advanced |  |  | which looks |
|  | Research in |  |  | insufficient to |
|  | Computer |  |  | execute with |
|  | Science and |  |  | long code as |
|  | Software |  |  | we know |
|  | Engineering |  |  | python is a |
|  |  |  |  | basic |
|  |  |  |  | language |
|  |  |  |  | with many |
|  |  |  |  | packages to |
|  |  |  |  | implement |
|  |  |  |  | code in some |
|  |  |  |  | lines . |
|  | **Text News** | Using Naïve | Statistical | It does not |
|  | **Classification** | Baye’s | characteristi | show how |
|  | **System using** | classifier and | cs of corpus | satisfactory |
| 2 | **Naïve Bayes** | language | and lexicon | results of text |
| **Classifier** | specific pre- | are | pre- |
|  |
|  | Publisher: An | processing | measured. | processing |
|  | International | techniques |  | module. |
|  | Journal of | for Punjabi |  |  |
|  | Engineering | language . |  |  |
|  | Sciences |  |  |  |
|  |  |  |  |  |
|  | **Sentiment** | Ttwitter data | we can | looks |
|  | **Analysis Using** | by using a | measure the | insufficient |
| 3 | **Naïve Bayes** | Naive | customers | classifier for |
|  | **Classifier** | Bayesian | opinions | news as it |
|  | Publisher: | algorithm. | and | just measure |
|  | International |  | perceptions. | the customers |
|  | Journal of |  |  | opinions in |
|  | Innovative |  |  | terms of |
|  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Technology and |  |  | positive or |
|  | Exploring |  |  | negative . |
|  | Engineering |  |  |  |
|  | (IJITEE) |  |  |  |
|  |  |  |  |  |
|  | **Comparison** | predict that | identify that | Unable to |
|  | **between** | whether the | which | identify |
|  | **Multinomial** | sentiment of | approach | clearly which |
|  | **and Bernoulli** | the news | between the | approach |
|  | **Naïve Bayes for** | article is | given two | between the |
| 4 | **Text** | positive or | approaches | given two |
|  | **Classification** | negative | perform | approaches |
|  | Publisher: | using the two | better for | perform |
|  | ICACTM | popular | the given | better for the |
|  |  | approaches | dataset. | given dataset |
|  |  | of Naïve |  |  |
|  |  | Bayes Text |  |  |
|  |  | Categorizatio |  |  |
|  |  | n. |  |  |
| 5 | **Document** | document | Results show | works on a |
|  | **Classification** | classification | that Naïve | limited data |
|  | **of Assamese** | using Naïve | Bayes is a | sets . |
|  | **Text Using** | bayes | good |  |
|  | **Naïve Bayes** | classifier. | classifiers. |  |
|  | **Approach** |  |  |  |
|  | Publisher: |  |  |  |
|  | International |  |  |  |
|  | Journal of |  |  |  |
|  | Computer |  |  |  |
|  | Trends and |  |  |  |
|  | Technology |  |  |  |
|  | (IJCTT) |  |  |  |
|  |  |  |  |  |
|  | **Some Effective** | naive Bayes | competing | we found a |
|  | **Techniques for** | in data | with state- | serious |
|  | **Naive Bayes** | mining tasks | of-the-art | problem in |
|  | **Text** |  | text | the parameter |
| 6 | **Classification** |  | classifiers | estimation |
|  | Publisher: IEEE |  | based on a | process, |
|  | TRANSACTIO |  | highly | which causes |
|  | NS ON |  | complex | poor results |
|  | KNOWLEDGE |  | learning | in text |
|  | AND DATA |  | method. | classification |
|  | ENGINEERING |  |  | domain. |
|  |  |  |  |  |

**4.METHODOLOGY**

Naïve Bayes Classification method of sentiment analysis is one of the widely used concept for predicting various types of sentiments in different contexts and situations. It is a probabilistic classifier and is mainly used when the size of the training set is less. In machine learning it is in family of sample probabilistic classifier based on Bayes theorem. The conditional probability that an event X occurs given the evidence Y is determined by Bayes rule by the (1).

P(X/Y) = P(X) P(Y/X) / P(Y) (1)

So for finding the sentiment the equation is transformed into the below (2). P(Sentiment/Sentence) = P(Sentiment)P(Sentence/Sentiment)/P(Sentence) (2) P(sentence/sentiment) is calculated as the product of P (token /sentiment) , which is formulated by the (3).

Count(Thistokeninclass)+1/Count(Alltokensinclass)+Count(Alltokens) (3) Here 1 and count of all tokens is called add one or Laplace smoothing Naive Bayes classifier is a general term which refers to conditional independence of each of the features in the model, while Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

**5.DATASET SOURCE:**

We have downloaded what we thought was the most appropriate dataset for news classification from Kaggle after some searching and studying all our options.

**5.IMPLEMENTATION**

1. **Importing the data:** The CSV file or the text file can be read and cleaned. In this case, thedata is already cleaned, so there is no need of preprocessing the data. A function to import the data that’s being read is defined.

Code:

%matplotlib

inline import

pandas as pd

titles = [] *# list of news titles*

categories = [] *# list of news categories*

labels = [] *# list of different categories (without repetitions)*

nlabels = 4 *# number of different categories*

lnews = [] *# list of dictionaries with two fields: one for the news and the other for its* *category*

**def** import\_data():

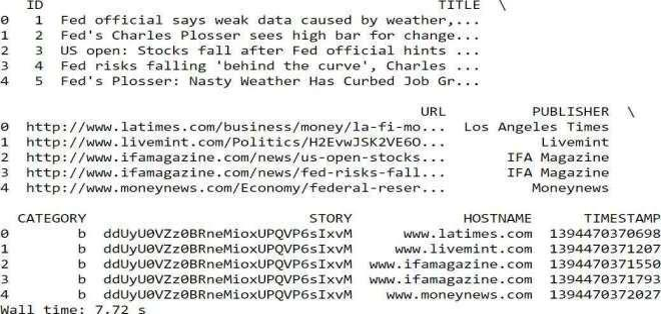
**global** titles, labels, categories

*# importing news aggregator data*

news = pd.read\_csv("uci-news-aggregator.csv")

*function 'head' shows the first 5 items in a column (orthe first 5 rows in the DataFrame)*  
  
*print(news.head()) categories = news['CATEGORY'] titles = news['TITLE']*  
  
*labels = sorted(list(set(categories)))*  
  
 *%***time** import\_data()

Output:



1. **Finding the number of categories and the headlines** that fall under each category. This helps in training and testing in later stages.

Input Code:

**from collections import** Counter

**def** count\_data(labels,categories):

c = Counter(categories)

cont = dict(c)

*# total number of news*

tot = sum(list(cont.values()))

d = {

"category" : labels,

"news" : [cont[l] **for** l **in** labels],

"percent" : [cont[l]/tot **for** l **in** labels]

}

print(pd.DataFrame(d))

print("total **\t**",tot)

**return** cont

cont = count\_data(labels,categories)

Output:

***Output:***

category news percent

* b 115967 0.274531
* e 152469 0.360943
* m 45639 0.108042
* t 108344 0.256485 total 422419

4) **plotting pie graphs** to get a better visual representation.

Input Code:

**import pylab as pl** *# useful for drawing graphics*

**def** categories\_pie\_plot(cont,tit):

**global** labels

sizes = [cont[l] **for** l **in** labels]

pl.pie(sizes, explode=(0, 0, 0, 0), labels=labels, autopct='**%1.1f%%**', shadow=**True**,

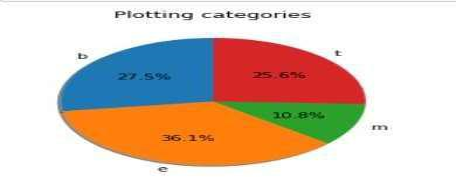
start angle=90)

pl.title(tit)

pl.show()

categories\_pie\_plot(cont,"Plotting categories")

Output:



**5) Splitting the data** into training and testing sets and use shuffle function to ensure its accuracyindependent of the frequency of occurrence of a particular category.

The training set evaluates predictive relationships and hence, categories. The testing set used to evaluate whether the discovered relationships hold and to assess the strength and utility of a predictive relationship.

Input Code:

**from sklearn.utils import** shuffle*# Shuffle arrays in a consistent way*

X\_train = []

y\_train = []

X\_test = []

y\_test = []

**def** split\_data():

**global** titles, categories

**global** X\_train, y\_train, X\_test, y\_test,labels

N = len(titles)

Ntrain = int(N \* 0.7)

titles, categories = shuffle(titles, categories, random\_state=0) *# Let's shuffle the data* X\_train = titles[:Ntrain]

y\_train = categories[:Ntrain]

X\_test = titles[Ntrain:]

y\_test = categories[Ntrain:]

%**time** split\_data()

cont2 = count\_data(labels,y\_train) *#analyzing the proportion of news categories in the* *training set*

categories\_pie\_plot(cont2,"Categories **% i**n training set")

Output:



**6.Training and testing:** In order to train and test the classifier, the first step should be totokenize and count the number of occurrence of each word that appears into the news' titles. Then the counters will be transformed to a TF-IDF representation. The last step creates the Multinomial Naive Bayes classifier.

Code:

**from sklearn.feature\_extraction.text import** CountVectorizer **from sklearn.feature\_extraction.text import** TfidfTransformer

**from sklearn.naive\_bayes import** MultinomialNB

**from sklearn.pipeline import** Pipeline

**from sklearn import** metrics

**import numpy as np**

**import pprint**

1. *lmats = [] # list of confussion matrix* nrows = nlabels  
     
      
   ncols = nlabels
2. *conf\_mat\_sum = np.zeros((nrows, ncols))*
3. *f1\_acum = [] # list of f1-score*

**def** train\_test():

**global** X\_train, y\_train, X\_test, y\_test, labels

*#lmats, \*

1. *conf\_mat\_sum, f1\_acum, ncategories* text\_clf = Pipeline([('vect', CountVectorizer()),  
     
    ('tfidf', TfidfTransformer()), ('clf', MultinomialNB()), ])  
      
   text\_clf = text\_clf.fit(X\_train, y\_train) predicted = text\_clf.predict(X\_test) **return** predicted  
     
    %**time** predicted = train\_test()  
     
    Output:Wall time: 27.1 s  
    metrics.accuracy\_score(y\_test, predicted)  
    Output:0.92380411281031516

print(metrics.classification\_report(y\_test, predicted, target\_names=labels))

Output

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| precision | recall f1-score |  | support |  |
| b | 0.90 | 0.91 | 0.90 | 34729 |
| e | 0.95 | 0.97 | 0.96 | 45625 |
| m | 0.97 | 0.85 | 0.90 | 13709 |
| t | 0.90 | 0.90 | 0.90 | 32663 |
| avg / total | 0.92 | 0.92 | 0.92 | 126726 |

**7)Confusion matrix** allows to detect if a classification algorithm is confusing two or moreclasses if you have an unequal number of observations in each class as in this case. An ideal classifier with 100% accuracy would produce a pure diagonal matrix which would have all the points predicted in their correct class.

mat = metrics.confusion\_matrix(y\_test, predicted,labels=labels)

cm = mat.astype('float') / mat.sum(axis=1)[:, np.newaxis]

Code:

import itertools

import matplotlib.pyplot as plt

def plot\_confusion\_matrix(cm, classes,

title='Confusion matrix',

cmap=plt.cm.Blues):

"""

This function prints and plots the confusion matrix.

"""

plt.imshow(cm, interpolation='nearest', cmap=cmap)

plt.title(title)

plt.colorbar()

tick\_marks = np.arange(len(classes))

plt.xticks(tick\_marks, classes, rotation=45)

plt.yticks(tick\_marks, classes)

thresh = cm.max() / 2.

for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):

plt.text(j, i, '{:5.2f}'.format(cm[i, j]),

horizontalalignment="center",

color="white" if cm[i, j] > thresh else "black")

plt.tight\_layout()

plt.ylabel('True label')

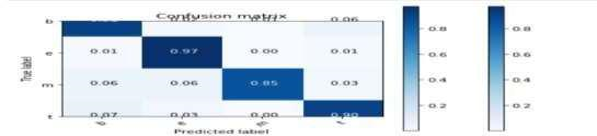
plt.xlabel('Predicted label')

plt.colorbar()

plt.show()

plot\_confusion\_matrix(cm, labels, title='Confusion matrix')

Output:



**def** resume\_data(labels,y\_train,f1s):

c = Counter(y\_train)

cont = dict(c)

tot = sum(list(cont.values()))

nlabels = len(labels)

d = {

"category" : [labels[i] **for** i **in** range(nlabels)],

"percent" : [cont[labels[i]]/tot **for** i **in** range(nlabels)],

"f1-score" : [f1s[i] **for** i **in** range(nlabels)]

}

print(pd.DataFrame(d))

print("total **\t**",tot)

**return** cont

f1s = metrics.f1\_score(y\_test, predicted, labels=labels, average=**None**)

cont3 = resume\_data(labels,y\_train,f1s)

***Output:***

category f1-score percent

0)b 0.903839 0.274738

1)e 0.959225 0.361334

2)m 0.902814 0.10798

3) t 0.903314 0.255945

total 295693

**6.RESULT**

Results show a good accuracy (0.9238) with a good average level for precision, recall and f1-score (0.92).

Analyzing these results by category, results are more accurate for the entertainment category ('e') with 0.96 for f1-score, 0.97 for recall and 0.95 for precision. Best result for prediction corresponds to health category ('m') with 0.97, but with a recall of 0.85.

**7.REFERENCES:**

1) Jasneet Kaur, Seema Bhagla, News Classification Using Naïve Baye’s Classifier, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 4, April 2016.

2) Shruti Bajaj Mangal, Dr. Vishal Goyal, Text News Classification System using Naïve Bayes Classifier, An International Journal of Engineering Sciences, Issue December 2014, Vol. 3 ISSN: 2229- 6913 .

3) Kavya Suppala, Narasinga Rao , Sentiment Analysis Using Naïve Bayes

Classifier, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.

4) G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, United Kingdom, 2019, pp. 593-596.

doi: 10.1109/ICACTM.2019.8776800

5) Moromi Gogoi 1 , Shikhar Kumar Sarma 2, Document Classification of Assamese Text Using Naïve Bayes Approach, International Journal of Computer Trends and Technology (IJCTT) – volume 30 Number 4

– December 2015.

6) Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, Some Effective Techniques for Naive Bayes Text Classification, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 11, NOVEMBER 2006.

7) Muhammad Abbas, Kamran Ali, Abdul Jamali, Multinomial Naive Bayes Classification Model for Sentiment Analysis, DOI: 10.13140/RG.2.2.30021.40169, March 2019

8) Ashraf Kibriya, Bernhard Pfahringer, Geoffrey Holmes, Multinomial naive Bayes for text categorization revisited, January 2004

9) R.Mohana, S.Sumathi, Document classification using Multinomial Naïve Bayesian Classifier, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 5, May 2014

10) N. Jindal and B. Liu, “Identifying Comparative Sentences in Text Documents”, Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 06), pp. 244-251, 2006.

11) A. McCallum and K. Nigam ”A comparison of event models for naive bayes text classification”.

12) Irina Rish. An empirical study of the naive bayes classifier. In IJCAI2001 workshop on empirical methods in artificial intelligence, pages 41–46, 2001.

13) K Raghuveer and K. N. Murthy ”Text categorization in indian languages using machine learning approaches” Department of Computer and Information Sciences,University of Hyderabad,Hyderabad.

14) Nidhi and V. Gupta, December-2012 ”Domain based classification of punjabi text documents using ontology and hybrid based approach”, Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING, pp. 109–122

15) J. D. Brutlag and C. Meek, ”Challenges of the email domain for text classification”, Microsoft Research, Redmond, WA, 98052 USA.

16) Rudy Prabowo, Mike Thelwall," Sentiment analysis: A combined approach ", Journal of Informetrics Volume 3, Issue 2, April 2009, pp 143–157.

17) Chowdhury Mofizur, Ferdous Ahmed Sohel et.al. “Text Classification using the Concept of Association Rule of Data Mining,” 2010. pp. 234-241

18) Ning Zhong, Yuefeng Li, and Sheng-Tang Wu “Effective Pattern Discovery for Text Mining” in IEEE transaction, vol. 24, January 2012.

19) S. Scott and S. Matwin, “Feature Engineering for Text Classification,” Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 379- 388, 1999.

20) RiniWongso, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli Rudy, "News Article Text Classification in Indonesian Language", Procedia Computer Science Volume 116, 2017, Pages 137-143