Application Research of Naive Bayes Algorithm Based on DIKW in Weather Website

Chaoning Li, Liang Chen, Shenghong Wu, Yunyin Mo and Liying Chen

June 23, 2021

Application Research of Naive Bayes Algorithm Based on DIKW in Weather Website

Chaoning Li[1],Liang Chen[1,2],Shenghong Wu[1],Yunyin Mo[1],Liying Chen[1]

1.Hainan Province Meteorological Service Center, Haikou Hainan 570100, China;

2.South China Sea Key Laboratory of Meteorological Disaster Prevention and Mitigation, Haikou Hainan 570203, China.

Abstract

The generation and update of weather website tourism products provide tourists with a reference for the weather conditions of the destinations. However, due to various reasons, they cannot be updated on time. Manual monitoring is required to update. If there is no update, manual update is required, which undoubtedly increases the burden on business personnel. The intelligent age needs to find a time-saving and labor-saving solution to solve this problem. The DIKW (Data Information Knowledge Wisdom) model is the most basic model in the research of information management, knowledge management, etc. The main content of the weather website is the collection of historical data, the summary storage of information, the knowledge processed by machine learning, and the wisdom of decision-making applications. The Naive Bayes Classification Algorithm (NBC) is widely used because of its high classification accuracy and simple model. For this reason, the model is combined with machine learning algorithms and imported into actual meteorological application research, and the NBC algorithm is combined with the weather website. One-month historical update data mining calculates the prior probability, and then the classification result is calculated according to the Python program to capture the data of the day. By recording the future 16 sample data sets using this model for calculation and analysis, 15 pieces of data conform to the results of the model calculation and classification, and the accuracy rate reaches 93.7%. The results show that the higher accuracy of the algorithm classification forecast can promptly remind business personnel, and better guarantee the timely update of tourism products, thereby improving the work efficiency of business personnel, and providing practical application reference value for smart weather service business automation.

Keywords:DIKW model;Naive Bayes;python;crawler;classification prediction;weather website

1 INTRODUCTION

In the era of rapid development of information modernization, busy weather data is being processed and transmitted every day. In order to ensure the timeliness, completeness, and accuracy of weather information data, it takes a lot of manpower to verify and correct the data, and transmit the corresponding data in time. The data is sent to government decision-making departments, disaster prevention, mitigation and flood prevention and other units. The weather

---

Fund Project: South China Sea Key Laboratory of Meteorological Disaster Prevention and Mitigation Open Fund (SCSF201907).

First author:Chaoning Li,(1990-), male, master, junior engineer, research direction meteorological information technology.

Corresponding author:Liang Chen, (1981-), male, senior engineer, research direction meteorological service and system development, Email: c_ray16@163.com.

service for the public is mainly to upload the processed weather data to the corresponding website platform on time, so that the public can easily and quickly check the weather conditions to arrange their own life, work and travel. However, in the process of transmitting weather data to the website platform due to various reasons, the data cannot be updated to the website platform in time, which requires human resources to monitor to ensure that the data is updated in time. This is undoubtedly an increase in the busy weather daily work. The workload of the salesman. How to use big data classification prediction technology to simplify the daily work of salespersons, the classification prediction algorithm of data mining obtains a classification result by training and analyzing a large amount of historical data. The naive Bayes algorithm model assumes that the attributes are independent of each other. It is best to use the naive Bayes algorithm when the attribute correlation is small. Therefore, the choice of the naive Bayes algorithm will affect the history of the website platform meteorological product data update. Data factors for data analysis and forecasting.

This article will combine the data information knowledge and wisdom (DIKW) model with smart weather application services. DIKW is a model of data, information, knowledge and wisdom, similar to a pyramidal hierarchical system, each of which Layers are endowed with certain characteristics than the next layer. By storing and processing the original meteorological data, useful information is obtained, and then the information is summarized, calculated, and deduced into knowledge to optimize the existing meteorological business products, which is effective The information can provide valuable decision-making services, and ultimately can effectively improve smart weather services.

2 DIKW MODEL

The DIKW system is a system of data, information, knowledge and wisdom. Data , Information , Knowledge , and Wisdom progress from the low-level to the high-level, and are interconnected with each other. Data is the most primitive material, which can be quantitative or qualitative, regardless of format. Information is a logical data flow after data processing. Knowledge is that humans make valuable information into knowledge through the understanding of data and information. Wisdom is human beings analyze information based on existing knowledge, and then make decisions about the future [9][10].

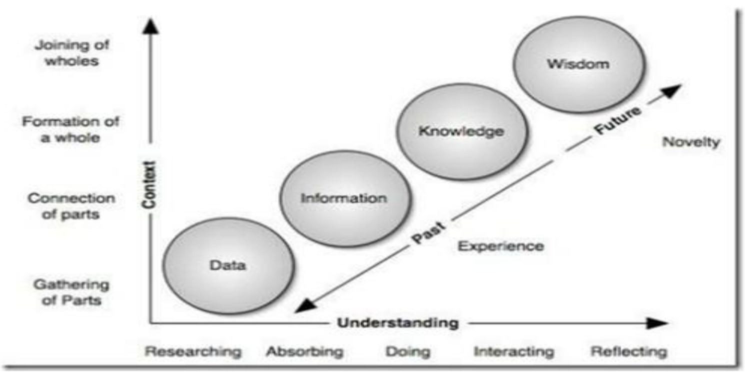

Figure 1: DIKW model

2.1 Meteorological application under DIKW model

Data can be structured or unstructured. It can be abstractly represented by collecting the attributes, quantity and mutual influence of the data that affect the update of the weather website, so that the data can be stored, processed and used later. By filtering and summarizing the discrete historical data that affects the update of the weather website, the probability and conditional probability of its attributes are calculated, and valuable data stream information is formed after processing. Combine these processed information with existing machine learning classification technology knowledge (linear regression algorithm, k-nearest neighbor algorithm, decision tree algorithm, naive Bayes algorithm, support vector machine algorithm, etc.), because of the naive Bayes algorithm Compared with other algorithms, it has a smaller error classification rate and is suitable for discrete data. The algorithm performance is stable and robust. Therefore, the weather website update result is calculated by the naive Bayes classification algorithm. This part of the information makes it useful Knowledge to assist the meteorological application business. Wisdom is the result of accumulation of experience and knowledge. It is used on the basis of knowledge to help business people make decisions to achieve the expected results.

3 PRINCIPLES OF NAIVE BAYES CLASSIFICATION ALGORITHM

3.1 Naive Bayes Method

The naive Bayes method is based on the Bayes principle and uses the knowledge of probability statistics to classify the sample data set. Due to its solid mathematical foundation, the misjudgment rate of Bayesian classification algorithm is very low. The characteristic of Bayesian method is the combination of prior probability and posterior probability, which avoids the subjective bias of using only the prior probability, and also avoids the overfitting phenomenon of using sample information alone [1]. Naive Bayes Classification (NBC) is a method based on Bayes' theorem and assuming that the feature conditions are independent of each other. First, the prior probability is calculated through the given training set, and then the largest one is obtained through calculation and comparison. Probability.

3.2 Naive Bayes Classification Process

3.2.1 Bayes Theorem

P(A|B) represents the probability of event A occurring under the premise that event B has occurred. It is called the probability of event A under event B. Its formula is as follows:

$$P(A \mid B) = \frac{P(AB)}{P(B)} \qquad (1)$$

Based on conditional probability, we can get P(B|A) through P(A|B):

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \qquad (2)$$

Among them, A represents the attribute set, B represents the class variable, P(B) is the prior probability, P(A|B) is the class conditional probability of the sample attribute A relative to the class variable B, P(A) is the evidence factor, P (B|A) is the posterior probability, and the Bayesian classification model expresses the posterior probability through the prior probability P(B), the class conditional probability P(A|B) and the evidence P(A).

The denominator P(A) in the above formula (2) can be decomposed into:

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i) \qquad (3)$$

3.2.2 Naive Bayes Classification Model

The Bayesian classification method will have problems such as sparse samples, discrete samples, and complex calculations in practical applications. In response to such problems, experts and scholars have proposed the assumption of independence on the conditional probability distribution of attributes, that is, by calculating the order of the probability values of various variables It is not necessary to use a completely accurate probability value to calculate the optimal value; secondly, the dependency between attributes sometimes has the same influence on all categories, and sometimes the influence brought by this dependency can offset each other, so The application of naive Bayes classifier can often get better and more accurate results [2][3].

Naive Bayes is an algorithm that relies on the Bayes rule, assuming that the conditions of each attribute are independent; the attribute vector set D={X1, X2,..., Xn}, the class vector set C={C1, C2 }, C1 is expressed as a positive category, C2 is expressed as a negative category, as in formula (4).

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)} \qquad (4)$$

For the distinction of distinguishing attribute classes, P(d) are all the same, that is, we get:

$$C = \operatorname{argmax} P(c \mid d), c \in C \qquad (5)$$

$$C = \operatorname{argmax} \frac{P(d \mid c)P(c)}{P(d)}, c \in C \qquad (6)$$

In formula (6), the denominators are all the same constant, and the comparison is negligible, that is, formula (7):

$$C = \operatorname{argmax} P(d \mid c)P(c), c \in C \qquad (7)$$

The attribute class can be expressed as:

$$C = \operatorname{argmax} P(x_1, x_2, ..., x_n \mid c)P(c), c \in C \qquad (8)$$

$$C = \operatorname{argmax} P(c)\prod_{i=1}^{n} P(x_i \mid c), c \in C \qquad (9)$$

As in formula (9), the comparison of the judgement attribute category is the comparison of the maximum value of the product of the prior probability.

4 EXPERIMENT AND ANALYSIS

4.1 Algorithm implementation

In the implementation of the naive Bayes algorithm, the classification results can be divided into whether the travel products of Hainan Weather Net are updated in time or not. Therefore, the product update results can be classified with 0, 1 classification problems. . The specific algorithm training application flowchart is as follows:

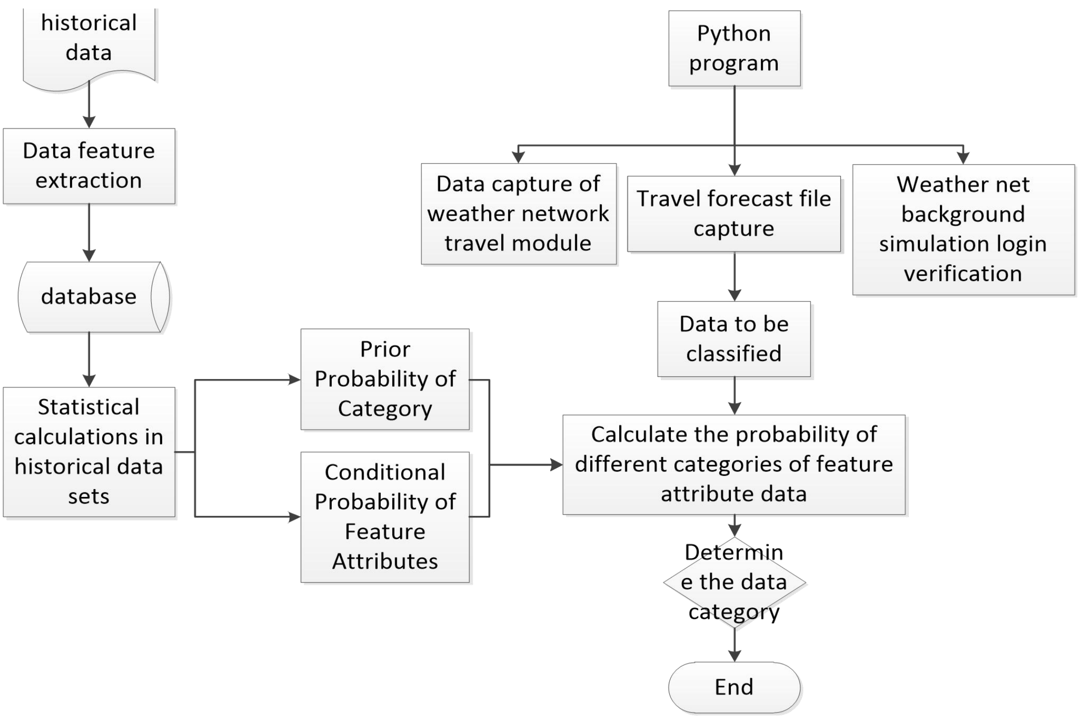

Figure 2:Naive Bayes algorithm classification process

As shown in Figure 1, during the training of the Naive Bayes algorithm, the recent historical data is first extracted into the database for processing, and the class conditional probability of each attribute and the prior probability of different categories are calculated statistically. The Python program passes three functional modules The latest data to be classified is automatically captured, and the posterior probability is calculated by the algorithm according to the currently acquired attribute feature value, and the maximum posterior probability is obtained by comparison, thereby obtaining the classification result.

4.2 Data preparation

The classification data source of this application comes from the data update record of the travel weather module of Hainan Weather Website in the past month. Data collection is carried out for the data update of the travel weather module at 19:00 every day. The training sample includes 31 data, and the attribute is a travel forecast file. The update status, the login status of the weather website background, and the update status of the weather website travel module are shown in Table 1 below.

Table 1:31 training data sets of tourism weather module

No	Update status of travel forecast documents	Weather net background login situation	Update status of weather net tourism module
1	File generation 1	Login normal 1	Data update 1
2	File generation 1	Login exception 0	Data not update 0
3	File not generation 0	Login normal 1	Data not update 0
4	File generation 1	Login normal 1	Data update 1
5	File generation 1	Login normal 1	Data update 1
6	File generation 1	Login exception 0	Data update 1
7	File generation 1	Login normal 1	Data update 1
8	File generation 1	Login exception 0	Data not update 0
9	File generation 1	Login normal 1	Data update 1
10	File generation 1	Login exception 0	Data not update 0
11	File generation 1	Login normal 1	Data update 1
12	File generation 1	Login normal 1	Data update 1
13	File not generation 0	Login normal 1	Data not update 0
14	File generation 1	Login normal 1	Data update 1
15	File generation 1	Login exception 0	Data not update 0
16	File generation 1	Login normal 1	Data update 1
17	File generation 1	Login normal 1	Data update 1
18	File generation 1	Login normal 1	Data update 1
19	File generation 1	Login normal 1	Data not update 0
20	File not generation 0	Login exception 0	Data not update 0
21	File generation 1	Login normal 1	Data update 1
22	File generation 1	Login normal 1	Data update 1
23	File generation 1	Login normal 1	Data not update 0
24	File generation 1	Login normal 1	Data update 1
25	File generation 1	Login normal 1	Data update 1
26	File not generation 0	Login normal 1	Data not update 0
27	File generation 1	Login normal 1	Data update 1
28	File generation 1	Login normal 1	Data update 1
29	File not generation 0	Login normal 1	Data not update 0
30	File generation 1	Login normal 1	Data update 1
31	File generation 1	Login normal 1	Data update 1

Table 2: Category cj and Sample statistics of attribute xi under cj condition

Update status of travel forecast documents		Weather net background login situation		Update status of weather net tourism module	
generate	not generate	normal	abnormal		
20	0	19	1	update	20
6	5	6	5	Not update	11

Table 3: Types of conditional probability P(xi|cj) and prior probability P(cj)

Update status of travel forecast documents		Weather net background login situation		Update status of weather net tourism module	
generate	not generate	normal	abnormal		
1.00	0.00	0.95	0.05	update	0.65
0.55	0.45	0.55	0.45	Not update	0.35

4.3 Application of Naive Bayes Model

Now the data obtained one day is X={file generation, abnormal login}, predict whether the travel data of that day will update the classification situation. According to Bayesian formula (7)(9):

$$P(c_{\text{update}}|X)=P(X\mid c_{\text{update}})P(c_{\text{update}})=\prod_{i=1}^{n}P(x_i\mid c_{\text{update}})*P(c_{\text{update}})$$

$$=P(x_{\text{file generation}}|c_{\text{update}})*P(x_{\text{login exception}}|c_{\text{update}})*P(c_{\text{update}})$$

$$=20/20*1/20*20/31=0.0322$$

$$P(c_{\text{not update}}|X)=P(X\mid c_{\text{not update}})P(c_{\text{not update}})=\prod_{i=1}^{n}P(x_i|c_{\text{not update}})*P(c_{\text{not update}})$$

$$=P(x_{\text{file generation}}|c_{\text{not update}})*P(x_{\text{login exception}}|c_{\text{not update}})*P(c_{\text{not update}})$$

$$=6/11*5/11*11/31=0.0879$$

Therefore, $P(c_{cap}\mid X)=\max(0.0322,0.0879)=0.0879$, it is predicted that the classification of tourism data for that day will not be updated.

When calculating the conditional probability of the sample, we found that when the i-th attribute value x_i is included, its conditional probability value is 0, if the attribute value of the sample to be estimated is x_i (the conditional probability value is 0) , Then the calculation result of the entire Bayesian formula will be 0. For example, we set the sample value as X={file not generated, login is normal}, and the value of $P(x_{\text{file generation}}|c_{\text{update}})$ is 0 in the calculation of the posterior probability. In order to calculate the conditional probability more accurately, Laplacian correction is introduced to solve this problem. The meaning is to add a prior distribution, corresponding to the case where the conjugate prior parameter takes 1, and expand the DC actual observation categories. The total number of categories in the training set is represented by N; the number of possible values of the Di attribute is represented by Ni, so the revised conditional probability calculation formula is:

$$P(x_{\text{i}}|c) = \frac{D_{c,xi} + 1}{D_c + Ni} \qquad (10)$$

The conditional probability corrected by Laplace is:

$$P(c_{\text{update}}|X)=P(X \mid c_{\text{update}})P(c_{\text{update}})=\prod_{i=1}^{n} P(x_i \mid c_{\text{update}}) * P(c_{\text{update}})$$
$$= P(x_{\text{file not generated}}|c_{\text{update}})*P(x_{\text{login normal}}|c_{\text{upda te}})*P(c_{\text{update}})$$
$$=1/22*19/20*20/31=0.0278$$

$$P(c_{\text{not update}}|X)=P(X \mid c_{\text{not update}})P(c_{\text{not update}})=\prod_{i=1}^{n} P(x_i \mid c_{\text{not update}}) * P(c_{\text{not update}})$$
$$= P(x_{\text{file not generated}}|c_{\text{not update}})*P(x_{\text{login normal}}|c_{\text{not update}})*P(c_{\text{not update}})$$
$$=5/11*6/11*11/31=0.0879$$

Therefore, $P(c_{cap} \mid X) = \max(0.0278, 0.0879) = 0.0879$,the result of the forecast classification is that the tourism data has not been updated, which is consistent with the previous calculation and classification results.

The following table 4 compares the update status of the travel forecast products recorded daily at 19:00 for the next 16 days and the calculation and classification results of the naive Bayes algorithm as shown in Table 4 (in the table, 1 means update, normal, 0 means not updated, abnormal). 11 of the results predicted by the algorithm under the condition that the travel forecast file is updated and the weather website back-end login is normal, are consistent with the actual weather website travel module update; the travel forecast file is not updated and the weather website back-end login is normal in the case of the algorithm prediction results One item is consistent with the fact that the tourism module of the actual weather website is not updated; the tourism forecast file is not updated and the weather website background login is abnormal. The result of the algorithm prediction is consistent with the situation of the tourism module of the actual weather website is not updated; the tourism forecast file is updated and In the case of abnormal weather website background login, one of the three predicted results of the algorithm is inconsistent with the actual weather website travel module update status. Analysis found that due to a short-term network failure in the business environment that day, the weather website background login verification program judged It was abnormal. After the network returned to normal, the weather website's travel module data was updated normally, which caused the algorithm prediction result to be inconsistent with the actual update of the weather website on that day. The prediction and classification results of 15 algorithms in 16 sample data are consistent with the actual situation, and the accuracy of model prediction reaches 93.7%.

Table 4 :Comparison of tourism product update results in the next 16 days

No	Update status of travel forecast documents	Weather net background login situation	Naive Bayes algorithm classification forecast update situation	The actual update of the weather net tourism module
1	1	1	1	1
2	1	1	1	1
3	0	1	0	0
4	1	1	1	1
5	1	0	0	0
6	1	1	1	1
7	0	0	0	0
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	***1***	***0***	***0***	***1***
12	1	1	1	1
13	1	1	1	1
14	1	0	0	0
15	1	1	1	1
16	1	1	1	1

4.4 Model evaluation

The Bayesian model algorithm is clear and easy to understand. The time and space overhead in the classification process is small, and the classification accuracy is high for discrete data. We conduct training and analysis by collecting data on different key factors that affect the update of weather website tourism products. The historical data is trained and calculated to get the classification result. Based on 31 training sample sets, the Naive Bayes classification model is used to classify and predict whether the weather website tourism products are updated. By recording the future 16 sample data sets, this model is used for calculation and analysis and the actual update situation is compared. The results show that 15 The results of data calculation and classification are consistent with the actual update situation, and the accuracy rate reaches 93.7%. The classification and prediction of the updates of other forecast products under the weather website, the result accuracy rate can also reach more than 90%, indicating that the model is effective. The predicted classification results through model calculation can achieve the expected results, and can more intuitively provide decision-making information to business personnel, thereby improving work efficiency and promoting the development of weather service business automation to a certain extent.

5 CONCLUSION

This paper discusses the algorithm principle of the naive Bayes model, combined with the DIKW model, uses the naive Bayes classifier to update the factors affecting the weather website tourism product data as attributes, and analyzes by collecting the historical data of the weather website as the training set. Extract the training data, delete redundant and edge-independent attributes, calculate its prior probability and class conditional probability, and obtain valuable information, and then calculate the classification prediction of the sample to be tested, and test the model with the next 16 sample data Evaluation, the experimental data results show that the prediction results of 15 data are consistent with the actual update situation, and the accuracy of the model prediction results and the actual update situation reached 93.7%, indicating that the naive Bayes classifier can achieve better classification results. Processing, obtain useful information, combine the knowledge of existing machine learning technology (naive Bayes classification algorithm), and then have the ability to predict the future. The disadvantage is that the training data and sample data are small, the model considers the factors that affect the update of the weather website tourism module is not comprehensive, and the model assumes that each attribute is independent of each other, so the classification effect is not good when the correlation between the attributes is strong. Using this model to classify and predict weather website product updates can effectively improve the work efficiency of business personnel, thereby reducing pressure on business personnel, and providing practical application reference value for smart weather service business automation.

REFERENCES

[1] Zhu Jun,Hu Wenbo. Recent Advances in Bayesian Machine Learning[J]. Journal of Computer Research and Development, 2015, 52(1): 16-26.

[2] WANG Leci,GAO Shichen,LIN Mengxiong,LI Zongxian. Naive Bayes classifier based on different probability density estimation methods[J]. China Mining Industry, 2018, 27(11):177-183.

[3] Ye Jin, Lin Shimin. Application of Bayesian network in subscriber churn analysis of wireless carriers[J]. Computer Applications,2005,25(3)673-675.

[4] Han Lina. Application of the Bayesian Classification Model in the Forecast of Student Achievementn[J]. Computer and Digital Engineering, 2018, 046(010):2039-2041,2056.

[5] Wang Bin.Research and implementation of spam filtering system based on Naive Bayes algorithm[J]. Electronic Design Engineering, 2018(17):171-174.

[6] Wang Jun. Extended Naive Bayes Classifier Based on Attribute Correlation Analysis[J]. Journal of Pingdingshan University, 2018, 33(05):70-74.

[7] Guo Xuncheng. Application Research of Naive Bayes Classification Algorithm[J]. Communication World, 2019, 26(01):247-248.

[8] Diao Haijun, Yin Zhao. A Data Prediction Based on Naive Bayes Classification Algorithm[J]. Radio and Television University Science and Technology, 2018, 277(04):5-7.

[9] Yucong Duan, Xiaobing Sun, Haoyang Che, Chunjie Cao, Zhao Li, Xiaoxian Yang.2019. Modeling Data, Information and Knowledge for Security Protection of Hybrid IoT and Edge Resources. IEEE Access 7: 99161-99176 (2019).

[10] Zhengyang Song, Yucong Duan, Shixiang Wan, Xiaobing Sun, Quan Zou, Honghao Gao, Donghai Zhu. Processing Optimization of Typed Resources with Synchronized Storage and Computation Adaptation in Fog Computing. Wirel. Commun. Mob. Comput. 2018: 3794175:1-3794175:13 (2018).

[11] Hermanto D T , Ziaurrahman M , Bianto M A , et al. Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier[J]. Journal of Physics Conference Series,2018,1140(1):012037.

[12] Sourav Kunal, Arijit Saha, Aman Varma, Vivek Tiwari.Textual Dissection of Live Twitter Reviews using Naive Bayes[J]. Procedia Computer Science, 2018,132:307-313.

[13] Findawati Y , Taurusta C , Widiaty I , et al. Teacher Performance Assesment Application using Naive Bayes Classifier Method[J]. IOP Conference Series Materials Science and Engineering, 2018, 384:012047.

[14] M Irfan,W Uriawan,O T Kurahman,M A Ramdhani,I A Dahlia. Comparison of Naive Bayes and K-Nearest Neighbor methods to predict divorce issues[J]. IOP Conference Series: Materials Science and Engineering,2018,434(1):012047.

[15] Deng X , Hou L , Wang F . Web advertisement detection using Naive Bayes[J]. Journal of Physics Conference, 2019, 1187(4):042023.

[16] Rahmadani S , Dongoran A , Zarlis M , et al. Comparison of Naive Bayes and Decision Tree on Feature Selection Using Genetic Algorithm for Classification Problem[J]. Journal of Physics: Conference Series, 2018, 978:012087.