Kristu Jayanti Journal of Computational Sciences Volume 3 2023, Pages 81-90

Advanced Naive Bayes Machine Learning System for Document Similarity Checking

M.Karthica[1], Dr.K.MeenakshiSundaram[2]

[1](Ph.D.Research Scholar, Department of Computer Science ,Erode Arts and Science College(Autonomous),Erode,Tamilnadu, India, karthica92@gmail.com)

[2](Associate Professor and Head, Department of Computer Science, Erode Arts and Science College(Autonomous),Erode, Tamilnadu,India,lecturerkms@yahoo.com)

Abstract

Content-based text analysis is recent development in machine learning that is clearly a technological advancement. In the digital age, everything is done quickly and instantly, giving business people greater insight. Classification is a crucial step in using machine learning in education to address these problems. Document writing appropriateness can be assessed by categorizing the topic-specific labeled training data. A content-based system is a tool that assists operators in locating content and overcoming the deluge of information. It assists in anticipating users' interests and provides recommendations based on the interest model of consumers. The evolution of collaborative filtering and the initial content-based recommender system are both continued, requires none of the user's checking appropriately.

Keywords: Content-based document classification advanced Naive Bayes method, citation analysis, document similarity, and natural language processing.

I. INTRODUCTION

Today's digital data development is proof that nearly all human endeavors have been enabled by modern technology. Rapid expansion of data is occurring. Digital data use is a contemporary trend that has a critical position for the requirements of all fields. The process of extracting data in order to derive knowledge and usable information is known as data mining. The information and knowledge gained during the mining process might be used for Market Analysis.

- Detection
- Prediction
- Classification
- Production control etc.

Fundamental data in text, numerical, ordinary, Boolean, and other forms has become more widely used as data mining has grown and developed. It grows swiftly, opening up machine learning methods. Text documents are divided into two categories: structured text data(18)and unstructured text data [2].In the digital age, social networking accounts for a large portion of data. Unstructured text data is the term used to describe this kind of data because it contains a large amount of information, such as hyperlinks, images, emotions, and naturally occurring languages like acronyms, uncommon terms, and the word "Alay."On the other hand, written materials including books, scientific articles, and abstracts are known as Text data before mining is done by processing from folding cases, tokenizing, stemming, and filtering [3],[5]. Additional pre-processing steps exist for managing data obtained from social media platforms [6]. Finding patterns, connections, and trends is the aim of text mining [4].One of the best aspects of data mining techniques is classification, which is simply the act of grouping.

A decision maker may find it difficult to make accurate decisions due to information overload caused by

the vast amount of information available on the internet. This is realized in real life when a user navigates a virtual shopping store with a lengthy list of items; It becomes more difficult to select items from a longer list. Content-based systems are software tools and methods that have been enhanced to help users find the items that interest them by anticipating their preferences or ratings. The concept is that I would get to know the users a little bit, so I would create a user profile based on their reviews of the products and suggest.

These days, content-based systems are a vital component of the majority of massive corporations such as Google, Face book, Amazon, and Netflix. They are used in a variety of applications, such as e-commerce [7], news [8], e-learning [1], entertainment, and healthcare. Numerous strategies have been developed to address the recommendation problem; conventional strategies include content- based filtering, collaborative filtering, and hybrid approaches..

These methods have serious issues, including low quality recommendations, scalability issues, low novelty and diversity, and high computational costs, even though they have been somewhat successful in offering relevant advice, especially after matrix factorization was introduced. Machine learning has also gained popularity in the field of content-based systems due to its advanced recommendation capabilities and ability to identify complex, non-linear relationships between users and items.

Deep learning models are typically computationally expensive, data-hungry, and non-interpretable. A multi-level search procedure has been used to gather pertinent papers. A content-based learning algorithm is used by content-based systems. A multi-level search engine was employed to locate relevant papers, and Reinforcement The term is "content-based learning system." There were about 44,000 papers found by this search. Following our initial screening stage, 1000 papers were selected from the initial 2000 articles that were found. In order to improve the dependability of our article collection, we have searched relevant libraries using the same keyword, such as the ACM Digital Library, IEEE Xplore, and Springer Link, up until the point where no more pertinent papers were found in the results.

II. RELATEDWORK

III.

Learning analytics has become the focal point of education research development because it provides valuable data and has the ability to accurately predict the study course Machine learning approaches are widely useful in the field of knowledge analytics. Paper [17] provides a comprehensive study of learning analytics by classification based on prediction algorithms, the RCV1 dataset, and factors given priority for prediction. Almost every aspect of education and the planning of the educational process make use of prediction models. Finding the links between the prerequisites, course characteristics, and subsequent students' success or failure is made easier with the help of the approach in [15]. [12] focuses on individual extraction troubles and also addresses overlap being relation appreciation and various entity relation extraction across sentences. A machine learning-based data recommendation system for e-learning, built on historical data, is represented by [1]. The authors demonstrate how recommendations for recently enrolled students are produced using data mining techniques like clustering and association rule algorithms [14]. Course prerequisites themselves maybe of scientific interest, even though the majority of studies, like [17], look into the relationships between them and students' academic success. This work focuses on developing recommendation systems that use machine learning algorithms to produce learning outcomes and prerequisites for relevant curricula.

Machine learning (ML) is a branch of artificial intelligence science that strives to mimic human brain function. A number of the ML presentations that can be finished involve absorbed learning, such as going to learning and control learning. such as Unsupervised Learning besides observational learning like as Reinforcement Learning. Currently machine learning is very helpful in various circles to analyse from

various case studies [11]. The interrelationships of machine learning types have been presented can be described as in Figure 1.1 below.

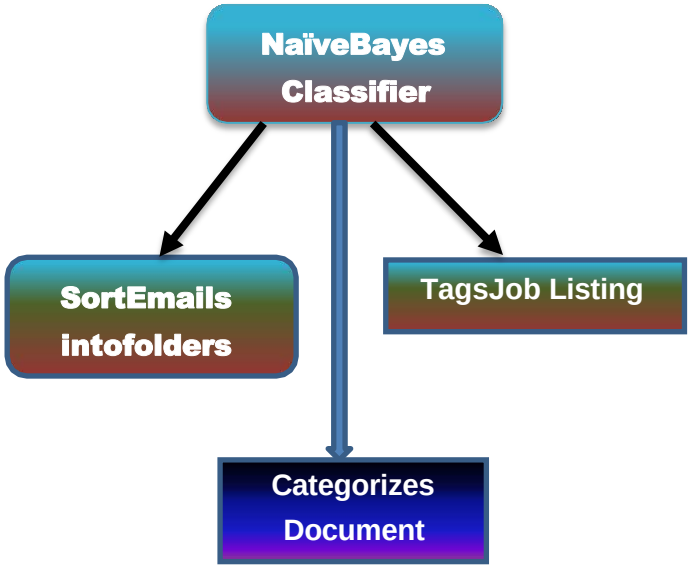


Fig1.1 MachineLearningandItsTypes

An algorithm established on Naive Bayes is a probabilistic organization algorithm. Established on strong autonomous expectations, it practices opportunity simulations. There is commonly no stimulus on faithfulness owing to autonomous expectations. As a result, they are considered naive. It is potential to train the Naive Bayes algorithm in supervised learning, dependent on the environment of the possibility classical. Naive Bayes prototypes encompass of a huge cube through the subsequent magnitudes:

- Name of the input field.
- Contingent on the input field type,the value range can be uninterrupted or discrete. By expending a Naive Bayes algorithm, uninterrupted fields get separated into discrete bins.
- Value of the target field.

The issue with the suggested approach is evaluating the written abstract that demonstrates the suggested method's degree of conformance. In order to trial the findings of summary writing of student-proposed work base on the identify, this research case will use both naïve based algorithm and python based machine learning to conduct abstract classification testing based on the before resolute proposed area labeling category. Filtering was not used in the pre-process method of the current method.

IV. METHODOLOGY

A study method called content analysis is used to discover particular terms, themes, or concepts in textual data that provide qualitative information. The presence, meanings, and relationships of specific keywords as well as words that are related to themes and concepts can all be examined by content analysis

researchers. The language employee din news article can be used by the user to draw conclusions about the messages contained within, the writer's audience, the culture and historical context of the text, hybrid texts, and link-based document representations. Links and texts provide important semantic information about the documents' discussion. Links show how a single document relates to the entire collection; the document text expresses the content as expressed by the author.

In table 1.1 gives the sample document counts and the composition of the abstract document. anticipating that useful information about a link's textual context will be revealed when assessing how contextually similar two documents are,

Table1.1.AbstractDocumentComposition

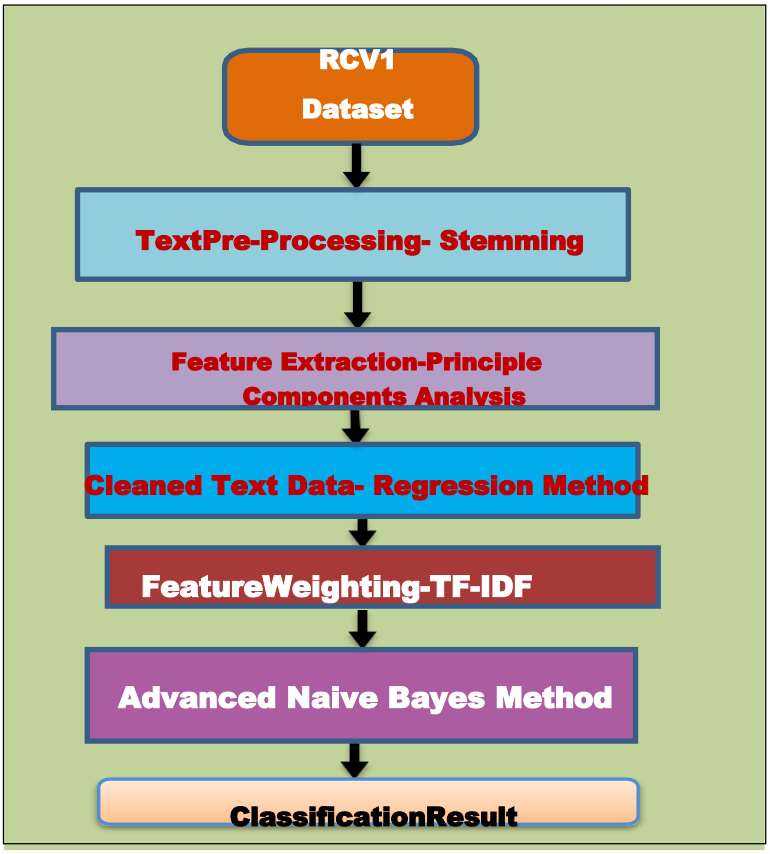
S.No	Topic	Number of Documents	Sample
1.	ResearchArticle	34	
2.	ImageProcessing	32	
3.	Data Mining	76	
4.	WeatherReport	24	
5.	Android Apps	24	

Segment level Document Similarity: To compute contextual document similarity, a straightforward method is to divide documents into semantic segments and calculate the similarity at the segment level. This kind of approach works for certain document types that have a standard format. Research papers can be divided into background, methodology, and conclusion sections. When the segmentation cannot be predefined or is found in non-standardized literature, this approach's proof of concept for segmentation is less straightforward. The difficulty lies in automatically dividing a coherent document into smaller parts. At its foundation, document segmentation is a task of sequence labeling. Traditionally, natural language text has been represented as a numerical vector using the vector space model or bag of words. With both approaches, documents are produced as sparse vector representations, where the values represent the frequency of a term in a document and the dimensions match the terms used in the document corpus. The values of the Term Frequency-Inverse Document Frequency(TF-IDF) are dependent upon the specificity of the terms. Sparse vector representations, when combined with cosine similarity, enable the effective computation of document similarities and are included in widely used information retrieval frameworks that have been effectively tested in content-based system research, such as Apache Lucene 1.

Content-based systems are commonly classified into four groups: knowledge-based, content-based, hybrid, and collaborative filtering. The primary focus of this work is on content-based recommender systems, which use the Advanced Naïve Bays Method to refer content referrers to any features that come from the recommended items. Concentrate on

theoretical works as the application field; these are also subject to related reviews. Kanakia et al. [4] developed a content-basedsystem for research papers using the Microsoft Academic Graph [4].

Fig.1.2ClassificationofusingAdvancedNaiveBayesMethod



The purpose of this research was to help the study program and the library guide students through the proposal process. That is, to give details on subjects that are popular right now and are frequently talked about annually. Figure 1.2 below provides a description of the research flow.

The consequences of classification tests on these areas with machine learning based Advanced Naive Bayes Method flowchart are displayed in figure 1.2 it can be explained. In this proposed method an entity relation feature extraction across sentences. Entity relations for feature extraction method based on sequence annotation by implementing a probabilistic model for text representation.

Data Collection:

The student proposal data collection library provided the information. Currently, the information is based on 2019– 2022. Data is collected in Excel format. Although it shouldn't be, the paragraph text continues from the subheading of the subsection. It is also advised to analyze the classification using the Advanced Naive Bayes method with the RCV1 dataset.

Text Pre-processing:

The pre-processing step in this case doesn't require too much labor because the text data used is taken from study findings or scientific publications in the form of abstracts. Eliminating Capitalization using the

lambda If char is not in the string, remove the purpose char for char in the text. Grammar. Eliminating Sequences Use the function re to get rid of numbers that are deemed unnecessary. Sub (text), '', '[0-9] +'. Situation transferable by extending the function lowercase sentence = sentence, associate all cultures with all lowercase. Decrease (). Tokenization Use the notch tokenizing function to divide each sentence into individual words. Stop phrase Remove conjunctions like "yang,""ke,""and,""di," and so forth by utilizing Corpus Words Indonesia's stop words feature. nltk.corpus.stopwords.words ('Indonesian') is the stop word. Rooting gets rid of suffixes.

Weighting:

The next step following pre-processing in this procedure was the weighing, which related to content-based systems. This is known as document frequency, and it involves calculating the rate of recurrence of occasion of words in each document. The function utilized in machine learning is taken from the sklearn library, and the equation used to compute this weighting is (1).Take Count Vectorizer from sklearn.feature_extraction.text.

Advanced Naïve Bayes Classifieris a gathering of organization algorithms based on Bayes Theorem. Each couple of type's existence classified is sovereign of every other. The data is separated in to two parts, namely quality matrix and there action vector.

- The attribute matrix contains all the vector of the RCV1 dataset in each vector consists of the value of needy features.
- The reaction and target vector (y) contains the value of class/ group variable for each row of feature matrix.

In order to finding this scikit-learn provides the utilities for the most common ways to extract numerical features from text content like,

- Tokenizing strings as well as generous an integer id for each feasible token, for instance by utilizing white- spaces and punctuation as token separators.

- Including the incident soft ok ensign every manuscript.
- Respectively discrete token existence occurrence is preserved as a feature.
  - The direction is entirely the indication frequencies for a known article are dignified a multi variants illustration.

Algorithm of Advanced Naive Bayes Algorithm

Step1: *Def separat_by_class(dataset):*

Step 2: *For i in range (len(dataset)): separated = dict()*

Vector = dataset[i] class_value=vector [-1]

Step 3: *separated[class_value] = list() separated[class_value].append(vector) return separated*

Step 4: *separated=separate_by_class(dataset) for label in separated: print(label)*

Step 5: *For row in separated [label]: print (row) def stdev(numbers):*

Step 6: *avg=mean(numbers)*

*variance=sum([(x-avg)**2for x in numbers]) / float(len(records)-1) return sqrt(variance)*

def mean(numbers):

Step 7: *return sum(numbers)/float(len(numbers))*

Step 8: *defstdev(numbers):*

avg=mean(numbers)

*variance=sum([(x-avg)**2forxinrecords])*

/float(len(numbers)-1) return sqrt(variance)

defsummarize_dataset(dataset):summaries=[(mean(column),stdev(column), len(column))

*for column in zip(*dataset)]*

del(summaries[-1]) return summaries

The calculation of Advanced Naive Bayes Method is

P(A|B)=P(B|A)*P(A)/P(B)

Where the probability is manipulative P(A|B) is called the posterior probability as well as the secondary probability ofthe event P(A) is called the prior.

Classification:

The Advanced Naïve Bayes Method (ANBM) is the classification technique employed. By using pre-existing libraries with function from sklearn.Naive_bayes multinomial introduce, machine learning can be achieved. Because there are eight target classes of suggested topics—research articles, image processing, datamining, weather reports, and android applications—this study uses a multinomial approach.

The proposed method used Advanced Naive Bayes method to filtering the content based system is in the pre-process process. So, the test in this text pre-processing will add a filtering process, to yield the precision, recall, and accuracy values of the method used.

V. RESULTS AND DISCUSSIONS

The content-based systems rely on the data use in the previously labeled abstract of data information proposed methods. Based on the data, the classes have been determined to cover major topics. A multinomial model using the Advanced Naive Bayes classifier method is used to test this abstract classification. There will be up to 200 documents in the article data in 2021–2023. Digital library performs automated citation indexing utilizing web content mining e- services include e- banking, search engines, accessible information management, collective networking, blog investigation, and personalization as well as suggestion systems. Web content mining prospered in removing the information from query lines as well as then matches correlated qualities. Using the Advanced Naive Bayes Classifier Multinomial method, 200 documents will be tested. The structure of training and test documents using the Weather Report and Android apps for the obtained training and check data is described in Table 1.2.

Table1.2Composition of training and test Documents

Data	Sum
Practice	126
Test	84

Regarding the recall results, certain class documents or topics—namely, image processing, research articles, and data mining—receive less favorable results. In order for the accuracy and recall ethics f1 score to still receive an

Average answer above80%.The precision results on the entire ANB Multinomial article obtain the standard values.

Table1.3.Test Results of Each Topic(in Percentage)

Topic	Precision	Recall	F1-Score
ResearchArticle	93	91	93
ImageProcessing	91	89	92
DataMining	86	85	84
WeatherReport	94	89	92
AndroidApps	90	91	94

In Table 1.3 Split results obtained using random state =1. The test results with the ANBM values are Research Article, Image processing, Data Mining, Weather Report and Android Apps.

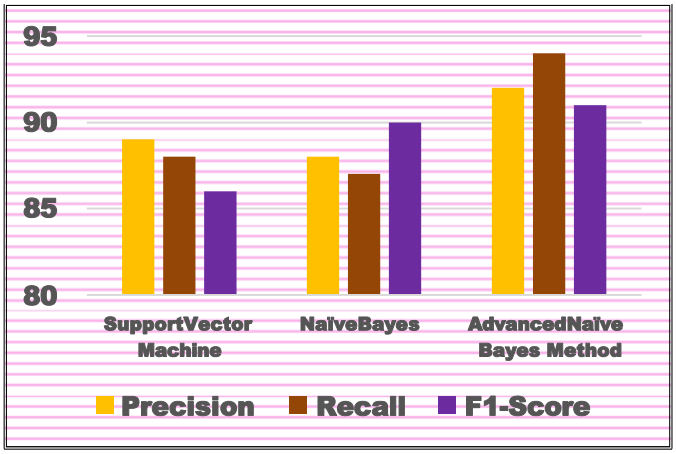


Fig1.2theComparison of ANBM for RCV1 Dataset

In Fig 1.2 explains the outcomes of using the Advanced Naive Bayes Method for text document similarity mining classification. The suggested Advanced Naïve Bayes Method yields superior results for precision, recall, andF1-Score metric values when compared to the current methods, such as support vector machines and Naive Bayes. According to the researcher's abstract study data, the precision value that has consequences is less real the more documents there are on the subject of data mining. There are only three when the data is unevenly composed, like in the case of e-commerce class documents. As a result, the fragmented trial is chosen with random state = 1 when exploited. If the random state expending is 42, then a document with a modest amount.

VI. CONCLUSION

In conclusion, the development of contextual metrics for document similarity could have a significant positive impact on literature document similarity. With the system that has been designed, users will be able to query documents and their relationships with each other in order to investigate document collections. Unlike current article similarity events, which only distinguish between similar and dissimilar documents and do not convey what makes two papers alike, the investigate compute will provide a context for the parallel. A machine learning system issued to determine the document similarity using the Advanced Naïve Bayes method. A recommended approach that outperforms the existing Naïve Bayes Algorithm is the Advanced Naïve Bayes Method.

REFERENCES

[1] Ahera, S.B., Lobo, L.M.R.J.: Combination of machine learning algorithmsfor recommendation ofcoursesin E-Learning System based on historical data. Knowl.-Based Syst. 51, 1–14 2013.

[2 AKR 19] Akromunnisa K and R.Hidayat,"Klasifikasi Dokumen Tugas Akhir (Skripsi) Menggunakan K-Nearest Neighbor," JISKA J. Inform. Sunan Kalijaga, vol. 4, no. 1, p. 69, May2019, doi: 10.14421/jiska.2019.41-07, 2019.

[3] Andrew Collins and Joeran Beel. 2019. Document Embeddings vs. Keyphrases vs. Terms: An Online Evaluation in Digital Library Recommender Systems. In ACM/IEEE Joint Conference on Digital Libraries (JCDL), pages 130–133, 2019.

[4] Anshul Kanakia, Zhihong Shen, Darrin Eide, and KuansanWang. 2019. A Scalable Hybrid Research Paper Recommender System for Microsoft Academic. In The World Wide Web Conference on - WWW "19, pages 2893–2899, New York, NewYork, USA. ACM Press, 2019.

[5] Asril H and I. Kamila, "Klasifikasi Dokumen Tugas Akhir Berbasis Text Mining menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor," p. 10, 2019.

[6] Atmaja D.M.U and R. Mandala, "AnalisaJudulSkripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor," IT Soc., vol. 4,no. 2, Aug. 2020, doi: 10.33021/itfs.v4i2.1182, 2020.

[7] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," J. Teknol. Inf., vol.3, no.2,p.179,Dec.2019,doi:10.36294/jurti.v3i2.1077,
2019.

[8] Feldman R and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge; New York: Cambridge University Press, 2007. Accessed:Feb. 20, 2021.

[9] Hidayatullah A.F and M. R. Ma"arif, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," p. 4, 2016.

[10] Kalokasari, D.H I. M. Shofi, and A. H. Setyaningrum, "Implementasi Algoritma Multinomial Naive Bayes Classifier Pada Sistem Klasifikasi Surat Keluar (Studi Kasus: DISKOMINFO Kabupaten Tangerang)," J. Tek. Inform., vol. 10, no. 2, pp. 109–118, Oct. 2017, doi: 10.15408/jti.v10i2.6199,
2017.

[11] Krol, Ed S etal.: Association between prerequisitesand academic success at a Canadian university"s pharmacy

program. Am. J. Pharm. Educ. 83(1) 2019.

[12] Liu, Q., Jia, X., Yang, W., Tu, F., Wu, L.: Research on entity relation extraction based on BiLSTM-CRF classical probability word problems. In: 13th International Conference on Education Technology and Computers. Association for Computing Machinery, pp. 62–68. NewYork, NY, USA 2021.

[13] Malte Schwarzer, Moritz Schubotz, NormanMeuschke, and Corinna Breitinger. 2016. Evaluating Link-based Recommendations for Wikipedia. Proceedings of the 16th ACM/IEEE Joint Conference on Digital Libraries (JCDL„16), pages 191–200, 2016.

[14] F. O. Reynaldi And N. Hikmah, "Implementasi Machine Learning Pada Sistem Pets Identification Menggunakan Python Berbasis Ubuntu," p. 6, 2020.

[15] Sama, R., Thamarai, L., Dr. Paul, P. Victer.: A survey on predictive models of learning analytics. Proc. Comput. Sci. 167, 37–46, 2020.

[17] O. Somantri, S. Wiyono, and D. Dairoh, "Metode K- Means untuk Optimasi Klasifikasi TemaTugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)," Sci. J. Inform., vol.3, no.1, pp. 34–45, Jun. 2016, doi: 10.15294/sji.v3i1.5845,
2016.

[18] Talbi, O., Chelik, N., Ouared, A., Ali, N.: Additive explanations for student fails detected from course prerequisites. In: International Conference of Women in Data Science, pp.1–7. Taif University (WiDSTaif),2021.