Shikshan Sanshodhan : Journal of Arts, Humanities and Social Sciences ISSN: 2581-6241 Volume - 3, Issue - 5, Sept-Oct – 2020
Bi-Monthly, Peer-Reviewed, Refereed, Indexed Journal Impact Factor: 3.589
Received on : 27/09/2020 Accepted on : 15/10/2020 Publication Date: 31/10/2020

भारतीय भाषाओं के पाठ वर्गीकरण पर किये गए शोध कार्यों के तकनीक एवं पद्धति का सर्वेक्षण

जावेद शेख़
शोधार्थी (सूचना एवं भाषा अभियांत्रिकी केंद्र)
महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालय, वर्धा 442001
Email - sjavvedalam1008@gmail.com

सारांश : पाठ वर्गीकरण एक प्रक्रिया है जिसमें डिजिटल दस्तावेजों को उनके श्रेणियों जैसे कि खेल पर्यटन, शिक्षा, प्रद्यौगिकी आदि के आधार पर पर संरचित तरीके से व्यवस्थित एवं प्रबंधित किया जाता है। तकनीकी के इस दौर में वर्ड वाइल्ड वेब पर बहुत सारे सूचनाएं डिजिटल दस्तावेज़, सम्मलेन सामग्री, सोशल मीडिया, ईमेल इत्यादि के रूप में असंचारित तरह से पड़े है। जिसका संगणक द्वारा उपयोग करना मुश्किल हो जाता है। इसको आसान बनाने के लिए पूर्व-प्रसंस्करण प्रणाली एवं कलन विधि की आवश्यकता है। पाठ खनन प्राकृतिक भाषा संसाधन का एक ऐसा अनुप्रयोग है जिससे डिजिटल दस्तावेजों से महत्वपूर्ण सूचना को निकाला जाता है। पाठ वर्गीकरण पाठ खनन के क्षेत्र में महत्वपूर्ण शोध विषय में से एक है। वर्त्तमान समय में इस क्षेत्र में बहुत तेजी से शोध हो रहे है। प्रस्तुत शोध पत्र में भारतीय भाषाओं के पाठ वर्गीकरण पर किये गए शोध कार्यों के तकनीक तथा पद्धति का सर्वेक्षण किया गया है।

मुख्य शब्द : प्राकृतिक भाषा संसाधन, कृत्रिम बुद्धि, भाषाविज्ञान, पाठ वर्गीकरण, डिजिटल दस्तावेज़।

१ प्रस्तावना :

प्राकृतिक भाषा संसाधन कृत्रिम बुद्धि का अनुप्रयुक्त क्षेत्र है जिसमें भाषाविज्ञान एवं कृत्रिम बुद्धि का समावेश है। कृत्रिम बुद्धि की शुरुआत 1950 में हुई। एलन ट्यूरिंग (Alan Turing) ने अपने शोध पत्र "Computing Machinery and Intelligence" [1950] के माध्यम से बताया की संगणक भी मानव बुद्धि की तरह कार्य कर सकता है। इसके उपरांत 1956 में जॉन मकार्थी (John McCarthy) ने एक सम्मलेन में इसे "artificial intelligence" (कृत्रिम बुद्धि) नाम दिया। जॉन मकार्थी (John McCarthy) को कृत्रिम बुद्धि का जनक माना जाता है। 1957 में नोम चौमस्की (Noam Chomsky) ने transformational-generative grammar सिद्धांत दिया जो प्राकृतिक भाषा संसाधन के दृष्टिकोण से बहुत महत्वपूर्ण साबित हुआ। प्राकृतिक भाषा संसाधन मानवी भाषाओं को संगणकीय प्रारूप के बीच संबंध स्थापित करता है। प्राकृतिक भाषाओँ को संगणकीय प्रारूप में स्थापित करने के लिए संगणकीय भाषाविज्ञान का भी प्रयोग किया जाता है संगणकीय भाषाविज्ञान कहीं न कहीं प्राकृतिक भाषा संसाधन से संबंध रखता है। सूचना प्रौद्योगिकी के क्षेत्र में क्रांति आने के बाद वर्ल्ड वाइड वेब पर बहुत सारे दस्तावेज डिजिटल रूप में उपलब्ध है, अभी भी लगभग 80% दस्तावेज असंगठित और अर्द्ध-संरचित रूप में है। आज के इस तकनिकी दौर में इन दस्तावेजों को संगठित रूप में करना अति आवश्यक है। इन अवश्यकाओं को देखते हुए इस क्षेत्र में शोध हो रहे हैं। डिजिटल दस्तावेज़ों को प्रोसेस करने के लिए पाठ खनन तकनीक का प्रयोग किया जा रहा है। पाठ खनन के कई अनुप्रयोग है जिसका उपयोग अगल-अगल कार्यों (जैसे सेंटीमेंट एनालिसिस, टॉपिक लेबलिंग, लैंग्वेज डिटेक्शन, पाठ वर्गीकरण) के लिए किया जाता है।

२ पाठ वर्गीकरण:

पाठ वर्गीकरण प्राकृतिक भाषा संसाधन का मौलिक कार्य है जिसमें डिजिटल दस्तावेज़ों को उनके श्रेणियों जैसे खेल, पर्यटन, शिक्षा आदि के आधार पर वर्गीकृत किया जाता है। पाठ वर्गीकरण, पाठ खनन का अनुप्रयुक्त क्षेत्र है। पाठ खनन का प्रयोग मुख्यतः पाठ के प्रबंधन एवं असंचरित पाठ से सूचनाओं को निकालने के लिए किया जाता है जिससे संगणक द्वारा इसका प्रयोग आसानी से किया जा सके। सर्वप्रथम David और उनके साथियों द्वारा 1994 में पाठ वर्गीकरण पर कार्य किया गया था। जिसमें इन्होने दो अल्गोरिथम का तुलना किया था।उसके बाद 1996 में Larkey एवं साथियों द्वारा पाठ वर्गीकरण पर कार्य

किया गया जिसका शीर्षक "Combining classifiers in text categorization" था। Joachims द्वारा 1998 में में पाठ वर्गीकरण पर कार्य किया जिसका शीर्षक "Text categorization with support vector machines: learning with many relevant features" जिसमें इन्होने svm अल्गोरिथम का प्रयोग किया था।

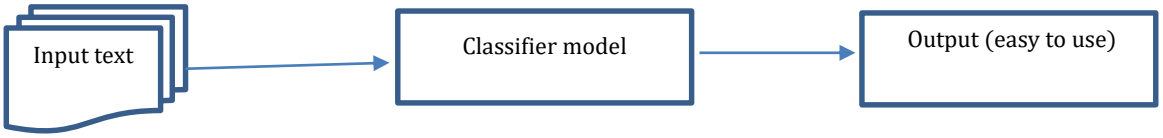

Figure no. 01 Simple block diagram of Text classification

पाठ वर्गीकरण को दो अलग-अलग तरीकों से किया जा सकता है

1) मैनुअल पाठ वर्गीकरण
2) स्वचालित पाठ वर्गीकरण

मैनुअल पाठ वर्गीकरण :

मैनुअल पाठ वर्गीकरण में पाठों का वर्गीकरण मनुष्य द्वारा पाठ को एनोटेट करके उनके श्रेणीय के आधार पर किया जाता है इस तकनीक में शुद्धता अधिक मिलती है। परन्तु समय ज्यादा लगता है।

स्वचालित पाठ वर्गीकरण :

स्वचालित वर्गीकरण में पाठ का वर्गीकरण संगणक द्वारा स्वचालित रूप से किया जाता है। इसके लिए कुछ नियम और कलन विधि का प्रयोग किया जाता है।

स्वचालित पाठ वर्गीकरण के लिए कई एप्रोच का प्रयोग किया जाता है। जिन्हें तीन समूहों में बांटा जा सकता है।

i) नियम आधारित एप्रोच
ii) मशीन लर्निंग आधारित एप्रोच
iii) हाइब्रिड एप्रोच

नियम आधारित अप्रोच :

नियम आधारित एप्रोच में भाषाई नियमों का प्रयोग करके पाठ को वर्गीकृत किया जाता है इस में पाठ को वर्गीकृत करने का एक तरीका ये हो सकता है यदि कोई समाचार पत्र है और उसको दो समूह में वर्गीकृत करना है तो पाठ को वर्गीकृत करने के लिए सबसे पहले संबंधित क्षेत्र के शब्दों का शब्दकोश तैयार करते हैं उसके बाद इनपुट के रूप में दिए गये पाठ में यह देखा जायेगा कि किस क्षेत्र का शब्द अधिक है पाठ को उसी क्षेत्र का माना जायेगा।

मशीन लर्निंग अप्रोच :

मशीन लर्निंग एक ऐसा एप्रोच है जिसमें किसी भाषाई नियमों की आवश्यकता नहीं पड़ती। इस एप्रोच में केवल डाटा को ट्रेन किया जाता है। उसी ट्रेन किए हुए डाटा से संगणक सिखाता है। और स्वचालित रूप से पाठ का वर्गीकरण करता है। मशीन लर्निंग के तीन मुख्यतः एप्रोच होते है।

1. Supervised machine learning
2. Unsupervised machine learning
3. Semi-supervised machine learning

सुपरवाइज्ड मशीन लीर्निंग एप्रोच (Supervised machine learning approach):

Supervised learning एक ऐसा प्रोसेस है जिसमें हम मशीन को तैयार करते समय labelled data का प्रयोग किया जाता है। Supervised learning में मॉडल को तैयार करते समय इनपुट और आउटपुट के अनुसार labelled data दिया जाता है।

अन-सुपरवाइज्ड मशीन लीर्निंग एप्रोच(Un-supervised machine learning approach):

unsupervised learning में मॉडल तैयार करते समय किसी भी प्रकार के labelled data कि जरूरत नहीं पड़ती है। unsupervised learning में मॉडल को इस तरह तैयार किया जाता है कि उसमें दिये unlabelled dataset में से जीतने भी एक समान के डेटा होते हैं उनका एक ग्रुप बन जाता है जिन्हें cluster कहा जाता है।

सेमी-सुपरवाइज्ड मशीन लीर्निंग एप्रोच (Semi-supervised machine learning approach):

semi-supervised learning, में मॉडल को तैयार करते समय labelled तथा unlabelled दोनों डेटा की आवश्यकता पड़ती है। कम मात्रा में labelled data की तथा ज्यादा मात्रा में unlabelled data की अवश्यकता

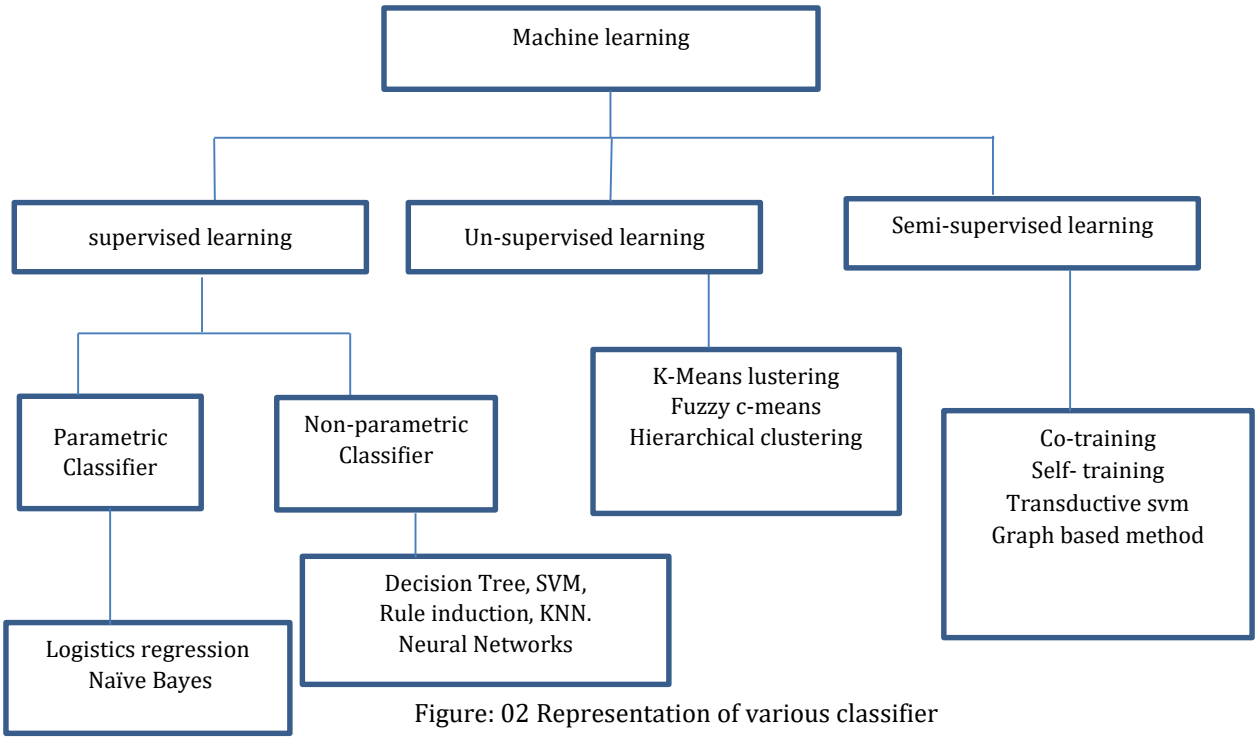

Figure: 02 Representation of various classifier

संकर आधारित एप्रोच:

इस एप्रोच में मशीन लर्निंग एवं नियम आधारित दोनों एप्रोच का प्रयोग किया जाता है। संकर आधारिक एप्रोच का प्रयोग इस लिए किया जाता है की परिणाम और शुद्धता को बढाया जा सके। जब मशीन लर्निंग एल्गोरिथ्म संदिग्धार्थता के कारण कुछ पाठ को वर्गीकृत नहीं कर पाता तो उसके लिए नियम आधारित एप्रोच का प्रयोग कर पाठ का वर्गीकरण किया जाता है। इसमें मशीन लर्निंग के मुकाबले अधिक शुद्धता मिलता है।

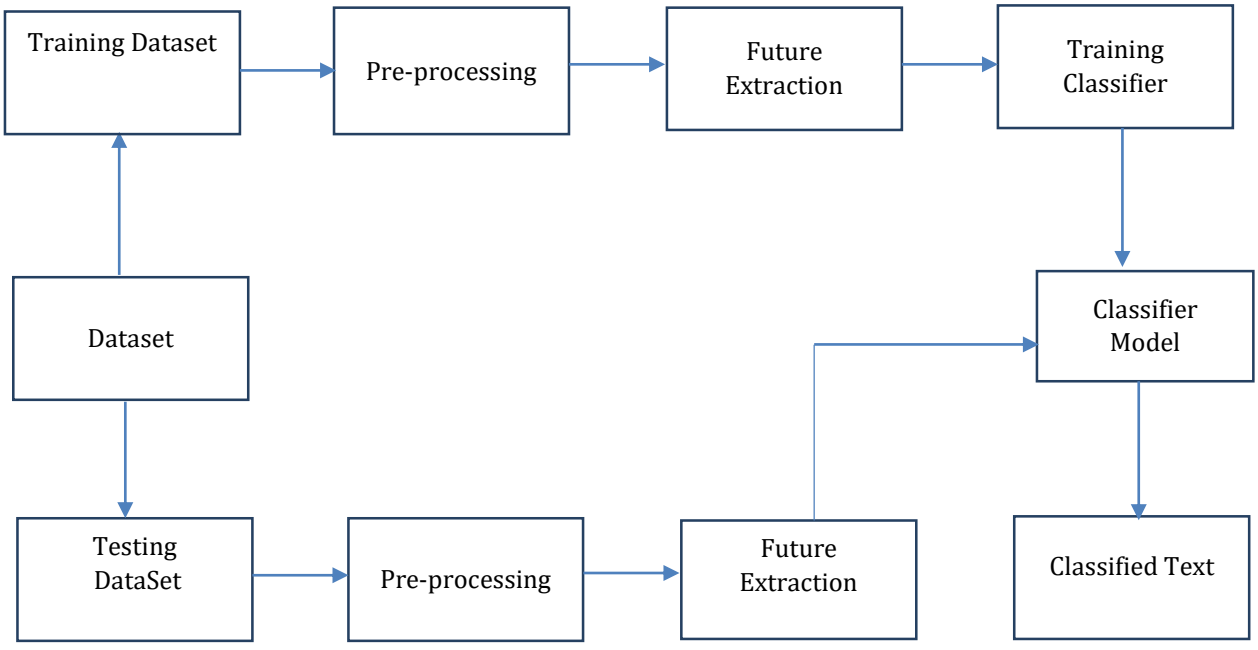

Figure: 03 Block Diagram of Text Classification

३ सर्वेक्षण :

- Narayana Swamy एवं M. Hanumanthappa द्वारा में लिखित शोध पत्र जिसका शीर्षक "Indian Language Text Representation and Categorization Using Supervised Learning Algorithm" में तीन भारतीय भाषाओं (कन्नड़, तेलेगु, तमिल) का वर्गीकरण किया गया है। इसके लिए इन्होने प्रत्येक भाषा के 100 दस्तावेजों को ट्रेन किया। और इनका परिक्षण तीन अल्गोरिथम Decision Tree Algorithm, Naive Bayes Algorithm, Nearest Neighbor Algorithm पर किया। जिसमें इन्हें Naïve Bayes gives Algorithm में 97.66%,Decision tree Algorithm में 97.33%, Nearest Neighbor Algorithm में 93% शुद्धता प्राप्त हुआ।
- U. Sree Krishnaetal द्वारा 2019 में लिखित शोध पत्र जिसका शीर्षक "Text Classification Using Fuzzy Neural Network" में लेखक ने पाठ वर्गीकरण के लिए फजी न्यूरल नेटवर्क एप्रोच का प्रयोग किया गया जिसमें उन्हें 95% शुद्धता प्राप्त हुआ।
- Farahadeebaandetal द्वारा लिखित जिसका शीर्षक "Urdu text Genre identification" है। इस शोध पत्र में उर्दू पाठ के वर्गीकरण के लिए SVM एप्रोच का प्रयोग किया गया है। जिसके लिए उन्होंने दो डाटा सेट (dataset) तैयार किया पहले डाटा सेट में 229 दस्तावेज़ों को ट्रेन किया और 56 पर परिक्षण किया तथा दूसरे डाटा सेट में 686 दस्तावेज़ों को ट्रेन किया और 160 का परीक्षण किया।
- Ali Abbas Raza और Ijaz Maliha ने अपने शोध पत्र "Urdu Text Classification" 2016 में बताया है कि उन्होंने उर्दू पाठ वर्गीकरण के लिए Naive Bayes और SVM मॉडल का परीक्षण 26067 उर्दू दस्तावेज़ों पर किया जिसमें इन्हें 93.34% शुद्धता मिला।
- Muhammad Usman और Saba Ayub द्वारा लिखित शोध पत्र जिसका शीर्षक "Urdu Text Classification using Majority Voting" के माध्यम से लेखक ने बताया है कि उन्होंने पाठ वर्गीकरण के लिए 2176 दस्तावेज़ों का संकलन किया और उसका tokenization किया गया जिससे 3078012 token मिला जिसमें कुल 120166 शब्द मिले। संकलित हुए डाटा पर पांच अल्गोरिथम Multinomial Naïve Bayes classifier, Bernoulli Naïve Bayes Classifier, Linear SVM,Random Forest Algorithm, Linear SGD Classifier का प्रयोग किया गया जिसमें Naïve Bayes classifier पर 87%,Bernoulli Naïve Bayes classifier पर 84%, Linear Random Forest Algorithm, Linear SGD Classifierपर 89%,Random Forest Algorithm पर 83% तथा Linear SGD Classifier 90% शुद्धता प्राप्त हुआ।
- Pooja Bolaj एवं Sharvari Govilkar द्वारा लिखित शोध पत्र जिसका शीर्षक "Text Classification for Marathi Documents using Supervised Learning Methods" के लेखक द्वारा बताया गया है कि उन्होंने मराठी दस्तावेज़ों के वर्गीकरण के लिए Naïve Bayes, Modified K Nearest Neighbor एवं Support Vector Machine अल्गोरिथम का प्रयोग किया गया है।
- Nidhiand Vishal Gupta द्वारा लिखित शोध पत्र जिसका शीर्षक "Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach" में पंजाबी पाठ का वर्गीकरण किया गया जिसके लिए Centroid Based Classifier, Naïve Bayes Classifier एवं Hybrid Approach का प्रयोग किया गया है। जिसमें Hybrid Approach की शुद्धता 80% तथा Centroid Based Classifier की 66% एवं Naïve Bayes Classifier की 57% शुद्धता प्राप्त हुई।
- Aishwarya Sahaniand el al द्वारा लिखित शोध पत्र जिसका शीर्षक "Automatic Text Categorization of Marathi Language Document" है जिसमें लेखकों द्वारा मराठी पाठ वर्गीकरण के लिए LINGO Algorithm का प्रयोग किया जिसमें उन्हें 95% की शुद्धता प्राप्त हुई।
- Jumi Sarmah and et al द्वारा लिखित शोध पत्र में जिसका शीर्षक "A Novel Approach for Document Classification using Assamese WordNet" में असमीज भाषा पे पाठ का वर्गीकरण किया गया है जिसके लिए Assamese WordNet का प्रयोग किया गया है जिसमें 90।27% शुद्धता प्राप्त हुई।
- K. N. Murthy द्वारा लिखित शोध पत्र में जिसका शीर्षक "Automatic Categorization of Telugu News Articles" है इस शोध पत्र में लिखने द्वारा बताया गया है उन्होंने तेलुगु समाचार पत्रों के वर्गीकरण Naïve Bayes Algorithm का प्रयोग किया जिसमें उन्हें 93% की शुद्धता प्राप्त हुई।

- Mansur Mand et al द्वारा लिखित शोध पत्र जिसका शीर्षक "Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper corpus" में लेखक द्वारा बंगला समाचार पत्र कार्पस के वर्गीकरण के लिए N-Gram Algorithm का प्रयोग किया गया है।
- S. Mohanty, and et al द्वारा प्रस्तुत शोधपत्र जिसका "Semantic Based Text Classification Using WordNets: Indian Language Perspective" में संस्कृत पाठ का वर्गीकरण wordnet का प्रयोग कर के किया गया है।

S.N. No.	Author	Title	Methodology	Output
1.	Narayana wamy, Hanumanthappa	Indian Language Text Representation and Categorization UsingSupervised Learning Algorithm	Decision Tree Algorithm, Naive Bayes Algorithm, Nearest Neighbor Algorithm	97.66% 97.33% 93%
2.	U. Sree Krishna, etal	Text Classification Using Fuzzy Neural Network	Fuzzy Neural Network	95%
3.	Abbas Raza Ali, Maliha Ijaz	Urdu Text Classification	NaiveBayes SVM	93ı34%
4.	Farahadeeba, etal	Urdu text Genre identification	SVM	
5.	Muhammad Usman	Urdu Text Classification using Majority Voting	Random Forest Algorithm, Linear SGD Classifier	83% 90%
6.	Pooja Bolaj, Sharvari Govilkar	Text Classification for Marathi Documents usingSupervised Learning Methods	Naïve Bayes, Modified K Nearest Neighbor	
7.	Nidhi, Vishal Gupta	Punjabi Text Classification using NaïveBayes, Centroid and Hybrid Approach	Hybrid, Centroid Based Classifier Naïve Bayes Classifier	80% 66% 57%
8.	Aishwarya Sahani, et al	Automatic Text Categorization of MarathiLanguage Document	LINGO Algorithm	95%
9.	Jumi Sarmah and et al	A Novel Approach for Document Classification using Assamese WordNet	Assamese WordNet	90.27%
10.	K. N. Murthy	Automatic Categorization of Telugu News Articles	Naïve Bayes Algorithm	93%
11.	Mansur M and et al	Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper corpus	N-Gram Algorithm	
12.	Menaka, S. Radha	A Text Classification using Keyword Extraction Technique	K Nearest Neighbor Naïve Bayes Decision tree	95% 87% 98%
13.	K Raghuveer and Kavi Narayana Murth	Text Categorization in Indian Languages using Machine Learning Approache	Naïve Bayes K Nearest Neighbor SVM	60.30%
14.	S. Mohanty, et al	Semantic Based Text Classification Using WordNets: Indian Language Perspective	using Sanskrit wordnet	

15.	Rajan et al	Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network	Artificial Neural network Space Vector Model	93.33% 90.33%

४ निष्कर्ष:

निष्कर्ष के तौर पर हम कह सकते हैं कि आज, पाठ वर्गीकरण की आवश्यकता बहुत बड़ी मात्रा में दस्तावेजों के कारण होती है जिसे हम दैनिक रूप से उपयोग में लाते हैं। डिजिटल दस्तावेज़ों की स्थापना के बाद से पाठ वर्गीकरण कृत्रिम बुद्धि तथा प्राकृतिक भाषा संसाधन के क्षेत्र में एक महत्वपूर्ण विषय बन चूका है। सामान्य रूप से, पाठ वर्गीकरण में विषय आधारित टेक्स्ट वर्गीकरण और टेक्स्ट शैली-आधारित वर्गीकरण शामिल होता है। कोई भी पाठ विशेष विषयों के आधार पर लिखे जाते है, उदाहरण के लिए: वैज्ञानिक, लेख, समाचार विवरण, चलचित्र समीक्षा, और विज्ञापन। विषय-आधारित पाठ वर्गीकरण दस्तावेज़ों को उनके विषयों के अनुसार वर्गीकृत करता है। डिजिटल दस्तावेज़ों विशेष रूप से बड़े पैमाने पर वेब पेज के अलावे दूसरे इलेक्ट्रॉनिक दस्तावेज़ों व पाठ जैसे ई-मेल, चर्चा समूह, विज्ञापन आदि का वर्गीकरण पाठ वर्गीकरण के द्वारा सरलता से कर सकते हैं। इस सर्वेक्षण के अध्ययन से यह पता चला की पाठ वर्गीकरण के लिए उपयोग में लायी जाने वाली अल्गोरिथम में सबसे ज्यादा शुद्धता Decision tree अल्गोरिथम की है और पाठ वर्गीकरण के लिए सबसे ज्यादा Naïve Bayes, K Nearest Neighbor एवं SVM अल्गोरिथम का प्रयोग किया जाता है। इस शोध अध्ययन से एक टूल का निर्माण किया जायेगा।

संदर्भ सूची:

1. Alan Turing, “Computing Machinery and Intelligence”, 1950,
2. Abbas Raza Ali, & Maliha Ijaz. (2009) Urdu Text Classification.ResearchGate.
3. Ayub Saba Usman Muhammad.(2016) . Urdu Text Classification using Majority Voting।International Journal of Advanced Computer Science and Applications, 273-265।
4. et al Krishna U. Sree(2019). Text Classification Using Fuzzy Neural Network International Journal of Recent Technology and Engineering, 198-193।
5. et al Mansur Munirul. (2000). Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus Academia, 31-25।
6. et al Sahani Aishwarya. (2016). Automatic Text Categorization of Marathi, International Journal of Computer Science and Information Technologies,, 2301-2297।
7. et al Sarmah Jumi. (2012) . A Novel Approach for Document Classification using Assamese WordNet. 6th International Global WordNet Confrence, 329-324।
8. Farah adeeba and et al.Urdu text Genre identification।Ethnologue Languge of the World, 14-9.
9. Govilkar Sharvari Bolaj pooja.(2016). Text Classification for Marathi Documents using Supervised Learning।ResearchgGate, 10-6।
10. Gupta Vishal Nidhi।Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach।Computer Science&Information Technology, 252-245।
11. Hanumanthappa M. Swamy M Narayana. (2013) . Indian Language Text Representation and Categorization Using।International Journal of Data Mining Techniques and Applications, 257-251।
12. Murthy Kavi Narayana.tomatic Categorization of Telugu News Articles।
13. Mohanty, S, Santi, P. K. Mishra, Ranjeeta, Mohapatra, R. N. and Sabyasachi Swain (2006), “Semantic Based Text Classification Using WordNets: Indian Language Perspective”, The Third InternationalWordnet Conference (GWC 06)। DOI=10।1।1।134।866।
14. Raghuveer. K, and Kavi Narayana Murthy, "Text Categorization inIndian Languages using Machine Learning Approaches." IICAI. 2007.
15. Menaka, S. Radha, “Text Classification using Keyword ExtractionTechnique,” vol। 3, no. 12, pp. 734–740, 2013.
16. RAJAN, K, RAMALINGAM, V, GANESAN, M।, PALANIVEL, S। AND PALANIAPPAN, B. (2009). Automatic Classification of Tamil documents using Vector Space Model andArtificial Neural

Network In: Expert Systems with Applications, Elsevier, Volume 36 Issue 8, DOI= 10।1016/j।eswa।2009।02।010.

17. Larkey, L. S. and Croft, W। B। "Combining classifiers in text categorization"। In Proceedings ofSIGIR-96, 19th ACM International Conference on Research and Development in InformationRetrieval (Zurich, CH, 1996), pp। 289–297 1996
18. Joachims, T. "Text categorization with support vector machines: learning with many relevantfeatures". In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz,DE), pp। 137–142 1998.
19. David D. Lewis and Marc Ringuette, "A comparison of two learning algorithms for textcategorization", Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis andInformation Retrieval, Las Vegas, US 1994.
20. https://www।abelard।org/turpap/turpap।php