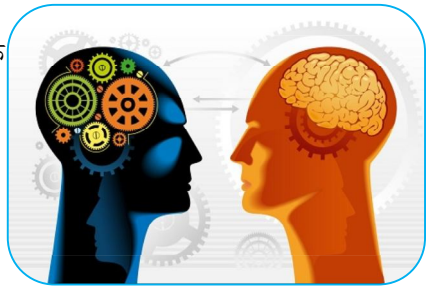
INDIAN STREAMS RESEARCH JOURNAL
ISSN NO : 2230-7850
IMPACT FACTOR : 5.1651 (UIF)
VOLUME - 12 | ISSUE - 5 | JUNE - 2022

"प्राकृतिक भाषा संसाधन के परिप्रेक्ष्य में विधिक पाठ सारांशीकरण: सर्वेक्षण "

अंजना किशनपुरी
शोधार्थी,सूचना एवं भाषा अभियांत्रिकी केंद्र,
महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालयमहाराष्ट्र। , वर्धा ,

I. शोध सारांश(Abstract)-



इस शोध-पत्र के माध्यम से भारतीय एवं विदेशी भाषाओं में विधिक डोमेन के लिए स्वचालित पाठ सारांशीकरण प्रणाली का एक व्यवस्थित सर्वेक्षण प्रस्तुत करने का प्रयास किया गया है। डिजिटल दुनिया के इस दौर में इंटरनेट(वेब) विभिन्न प्रकार की सूचनाओं से भरा हुआ है। बड़ी मात्रा में डेटा उत्पन्न किया जा रहा है और हर सेकंड लोगों और मशीनों द्वारा उपभोग किया जाता है। ऑनलाइन समाचार, कानूनी दस्तावेज, उत्पादों की समीक्षा, ऑनलाइन प्रश्न उत्तर मंच और विभिन्न सोशल मीडिया पोस्ट जैसे ट्वीट, फेसबुक पोस्ट, ब्लॉग, माइक्रो-ब्लॉग आदि डेटा उत्पन्न करने वाले स्रोतों के कुछ उदाहरण हैं। ये स्रोत प्रतिदिन अरबों दस्तावेज़ उत्पन्न कर रहे हैं। एक इंसान के लिए, बड़े दस्तावेज़ों को समझना और उसमें से महत्वपूर्ण जानकारी हासिल करना अक्सर एक श्रमसाध्य और समय लेने वाला कार्य होता है। जो सूचना के विशाल भाग का एक सुसंबध्द और सूचनात्मक संस्करण (लघु-सारांश) उत्पन्न करने की आवश्यकता को प्रेरित करता है। यह सुसंबध्द जानकारी प्रासंगिक जानकारीयों को जल्दी खोजने और समझने में मदद करती है। स्वचालित पाठ सारांशीकरण प्रणाली की आवश्यकता बढ़ रही है जो इनपुट पाठ दस्तावेज़ का पूर्ण और समझने योग्य सारांशित पाठ तैयार करेगा।पाठ सारांशीकरण प्राकृतिक भाषा संसाधन के केंद्रीय क्षेत्रों में से एक है। पाठ सारांशीकरण प्रणाली किसी भी भाषा में बड़े पाठ के सार्थक सारांश को सुगमबनाने में महत्वपूर्ण भूमिका निभाती है। इंटरनेट पर लगभग सभी भाषाओं में डेटा की व्यापक उपलब्धता है। इसलिए यहां कम समय में अधिक पढ़ने के लिए पाठ सारांशीकरण की आवश्यकता है और हम स्वयं भी तय कर सकते हैं कि इसे आगे पढ़ना है या नहीं। स्वचालित पाठ सारांशीकरण कई अनुप्रयोगों की आवश्यकता बन गया है, उदाहरण के लिए, बहुत छोटे समाचार, खोज इंजन, व्यवसाय विश्लेषण, विधिक आलेख सारांश इत्यादि। प्रस्तुत शोध-पत्र में हम विधिक पाठ सारांशीकरण के विभिन्न तरीकों एवं दृष्टिकोणों पर विस्तार से एक सर्वेक्षण प्रदान कर रहे हैं, जो वर्तमान परिदृश्य में अधिकतर उपयोग किए जाते हैं।

मूल शब्द (Keywords):- पाठ सारांशीकरण प्राकृतिक भाषा संसाधन। ,भावात्मक सारांशीकरण ,निष्कर्षण सारांशीकरण ,मशीन लर्निंग ,

II. परिचय -

डिजिटल दुनिया के इस दौर में इंटरनेट(वेब) विभिन्न प्रकार की सूचनाओं से भरा हुआ है। बड़ीमात्रा में डेटा उत्पन्न किया जा रहा है और हर सेकंड लोगों और मशीनों द्वारा उपभोग किया जाता है। ऑनलाइन समाचार, कानूनी दस्तावेज, उत्पादों की समीक्षा, ऑनलाइन प्रश्न उत्तर मंच और विभिन्न सोशल मीडिया पोस्ट जैसे ट्वीट, फेसबुक पोस्ट, ब्लॉग, माइक्रो-ब्लॉग आदि डेटा उत्पन्न करने वाले स्रोतों के कुछ उदाहरण हैं। ये स्रोत प्रतिदिन अरबों दस्तावेज़ उत्पन्न कर रहे हैं। एक इंसान के लिए, बड़े दस्तावेज़ों को समझना और उसमें से महत्वपूर्ण जानकारी हासिल करना अक्सर एक श्रमसाध्य और समय लेने वाला कार्य होता है। जो सूचना के विशाल भाग का एक सुसंबध्द और

सूचनात्मक संस्करण (लघु-सारांश) उत्पन्न करने की आवश्यकता को प्रेरित करता है। यह सुसंबध्द जानकारी प्रासंगिक जानकारीयों को जल्दी खोजने और समझने में मदद करती है। स्वचालित पाठ सारांशीकरणप्रणाली की आवश्यकता बढ़ रही है जो इनपुट पाठ दस्तावेज़ का पूर्ण और समझने योग्य सारांशित पाठ तैयार करेगा।

प्राकृतिक भाषा संसाधन(एनएलपी) पाठ्य डेटा को संसाधित करने के लिए उपयोग किए जाने वाले दृष्टिकोणों का एक समूह है। जिस संपूर्ण संसाधित पाठ पर कार्य किया जा रहा है उसे कार्पस कहा जाता है। इस तरह के संसाधित पाठ केकई उपयोग हैं जैसे- व्याकरण और वर्तनी की जांच करना, वाक् में अग्रिम वर्ण का ज्ञान, उपयोगकर्ता आधारित प्रश्न उत्तर सूचनाएं प्राप्त करना(प्रज्ञा(intelligence) इंटरफ़ेस, विधिक पाठ जिनकी चर्चा लेख में गई है), पाठ वर्गीकरण, सारांशीकरण इत्यादि।

विधिक डोमेन भाषा और शब्दार्थ तकनीकों के लिए एक आकर्षक डोमेन रहा है, क्योंकि यह शासन के लिए आवश्यक है और यह प्राकृतिक भाषा संसाधन में अनुसंधान में एक बड़ी भूमिका के रूप में उभर रहा है। हाल के शोध ने वैचारिक प्रश्नों कबीच एक सेतु बनाने की आवश्यकता पर प्रकाश डाला है, जैसे कि खनन(Mining) और तर्क(Reasoning) में विधिक व्याख्या की भूमिका, साथ ही कम्प्यूटेशनल और अभियांत्रिकीय चुनौतियां जैसे कि बड़े विधिक डेटा और नियामक अनुपालन की जटिलता को संभालना आदि **लिवियो(2019)**। इन दो उद्देश्यों पर प्रयासों को एकीकृत करने की दिशा में प्रगति की सुविधा के लिए, यूरोपीय संघ ने हाल ही में कई शोध परियोजनाओं को वित्त पोषित किया है जिनमें से 'MIREL: MIning and REasoning with Legal texts',http:// www.mirelproject.eu एक है।

वेब पर स्वतंत्र रूप से उपलब्ध बड़े डेटा को स्वचालित रूप से विश्लेषण, अनुक्रमण और समृद्ध करने के लिए ऐसे एनएलपी विधियों और अर्थीय(semantics) तकनीकों के विकास ने विधिक प्रणालियों की दक्षता, बोधगम्यता और निरंतरता में सुधार के लिए नये दृष्टिकोणों के निर्माण के अवसर पैदा किए हैं। उदाहरण के लिए, एक दृष्टिकोण वाक्यात्मक तत्वों को जोड़ना है, जैसे- संज्ञा,क्रिया और उपवाक्य, उनके शब्दार्थ वैसे ही किसी दिए गए डोमेन में जैसेव्यक्ति ,गुण और संबंध।

III. पाठ सारांशीकरण-

पाठ सारांशीकरण प्राकृतिक भाषा संसाधन का नया उभरता हुआ शोध का विषय है।पाठ सारांशीकरण text mining तथा प्राकृतिक भाषा संसाधन (NLP) तथा Natural Language Generation में महत्वपूर्ण भूमिका निभाता है ।

आज के दौर में भाषाविज्ञान के क्षेत्र में भाषा और मशीन के मध्य इस अंत:क्रियात्मकता पर और भी विचार विमर्श किए जा रहे हैं । प्रस्तुत शोध विषय सैद्धांतिक भाषाविज्ञान, भाषा अभियांत्रिकी, कृत्रिम बुद्धि और कंप्युटेशनल भाषाविज्ञान का समन्वित रूप है, जिसमें भाषाविज्ञान से संबंधित सैद्धांतिक नियमों को कंप्यूटर केद्वारा प्रोग्राम के माध्यम से जोड़ा गया है।

पाठ सारांशीकरण एक या अधिक पाठ से निर्मित होता है जो मूल पाठ में एक छोटे रूप में महत्वपूर्ण सूचना देता।हैपाठ सारांशीकरण के दो महत्वपूर्ण दृष्टिकोण हैं जैसे- भावात्मक दृष्टिकोण (Abstractive Approach),निष्कर्षण दृष्टिकोण (Extractive Approach).

भावात्मक दृष्टिकोण (Abstractive Approach)-इस तकनीक में पहले पूरे दस्तावेज़ को समझा जाता है और फिर मूल दस्तावेज़ से महत्वपूर्ण जानकारी निकालकर तथा उस जानकारी के उपयोग से नए वाक्यों को उत्पन्न करके एक पाठ दस्तावेज़ का संक्षिप्त(concise) और बोधप्रद (Instructive) संस्करण प्रदान किया जाता है। इस तकनीक के अंतर्गत् भाषावैज्ञानिक पद्धतियों का प्रयोग पाठ का गहन विश्लेषण करने के लिए किया जाता है। निष्कर्षण दृष्टिकोण (Extractive Approach)-इसके अंतगत् मूल टेक्स्ट में से महत्वपूर्ण शब्द वाक्य या वाक्यांस के उपसमुच्चय(subset) को निष्कर्षित (extract) कर क्रमबद्ध (concatenate) जोड़कर सारांश उत्पन्न किया जाता है। लेकिन इसकी यह सीमा है कि यह कभी भी बड़े दस्तावेज़ों के लिए मानव स्तर सारांश उत्पन्न नहीं कर सकता।

IV. साहित्य सर्वेक्षण-

1. Atefeh और Guy (2004) द्वारा विधिक आलेख के सारांशीकरण के लिए एक दृष्टिकोण प्रस्तावित किया है, जिसके द्वारा विधिक विशेषज्ञ को निर्णय के प्रमुख विचारों को निर्धारित करने में मदद प्राप्त होगी। इनका दृष्टिकोण पाठ की सुसंगतता और पठनीयता में सुधार के लिए एक तालिका शैली सारांश बनाने के लिए आलेख डिज़ाइन और उसकी विषयगत संरचनाओं की खोज पर आधारित है। इस दृष्टिकोण के साथ निर्मित उन्होंने एक प्रणाली के घटकों को प्रस्तुत किया है, जिसे LetSum कहा जाता है, जिसमें कार्यान्वयन और कुछ प्रारंभिक

मूल्यांकन परिणाम उपस्थित होते हैं। सारांश का निर्माणचार चरणों में किया जाता है- आलेख संरचनाओं का पता लगाने के लिए विषयगत विभाजन, महत्वहीन उद्धरणों और अवांछनीय पाठ को हटाने के लिए फ़िल्टरिंग, उम्मीदवार इकाईयों का चयन और तालिका शैली सारांश का उत्पादन।

2. Atefeh और Guy (2004) ने इस शोध-पत्र में न्यायिक निर्णय के स्वचालित सारांशीकरण के लिए एक नई पद्धति के विकास को प्रस्तुत किया है। LetSum (विधिक पाठ सारांशक) का वर्णन किया है, जो एक प्रोटोटाइप प्रणाली है। आलेख संरचना की पहचान करके और निर्णय खंडों के विषयों को निर्धारित करके स्रोत निर्णय में प्रासंगिक इकाइयों केनिष्कर्षण के आधार पर इन्होंने अपना दृष्टिकोण प्रस्तुत किया है। सारांश का निर्माण चार चरणों में किया जाता है चार विषयों- परिचय, संदर्भ, न्यायिक विश्लेषण और निष्कर्ष में निर्णय की विषयगत संरचना को निर्धारित किया गया है। फिर यह प्रत्येक विषय के लिए प्रासंगिक वाक्यों की पहचान करता है। इन्होंने साथ ही सांख्यिकीय पद्धति के साथ उत्पादित सारांशों के मूल्यांकन और न्यायिक निर्णय के आधार पर मानव मूल्यांकन पर भी चर्चाकी है। अन्य सारांशीकरण तकनीकों की तुलना में इस प्रणाली के परिणाम अच्छे प्रदर्शन का संकेत देते हैं।

3. Ben और Claire (2005) द्वारा विधिक आलेखों के सारांशीकरण के संदर्भ में वर्गीकरण प्रयोग प्रस्तावित किए गए हैं जिसके लिए इन्होंने अलंकारिक स्थिति एनोटेशन के अलावा विस्तृत भाषाई मार्कअप के साथ यूके हाउस ऑफ लॉर्ड्स के निर्णयों का एक नया कार्पस विकसित किया है। इन्होंने कई मशीन लर्निंग एल्गोरिथम की तुलना की है जिन्होंने पहले प्राकृतिक भाषा कार्यों पर अच्छा प्रदर्शन दिखाया है। इनमें से, सपोर्ट वेक्टर मशीन और अधिकतम एन्ट्रापी मॉडल उनके कार्य के लिए सबसे उपयुक्त साबित हुए हैं। इन्होंने व्यापक रूप सेउपलब्ध भाषावैज्ञानिक विश्लेषण टूल के आधार पर Cue phrase जानकारी को प्राप्त करने के लिए एक प्रभावी एवं सामान्य तरीका प्रस्तुत किया है। साथ ही वाक्य स्तरीय प्राकृतिक भाषा कार्य के लिए अनुक्रम मॉडलिंग दृष्टिकोण प्रस्तुत किया, जिसने आधारभूत वर्गीकरण पर प्रदर्शन में काफी सुधार किया।

4. Ben और Claire (2006) ने एक दस्तावेज़ से सबसे सारांश-योग्य वाक्यों के चयन के कार्य के लिए एवं कई प्रकार की विशेषताओं का उपयोग करके वाक्यों की अलंकारिक स्थिति का पूर्वानुमानकरने के कार्य के लिए मशीन लर्निंग की तकनीक का उपयोग करके प्रयोगात्मक परिणाम प्रस्तुत किए हैं। इन घटकों के लिए परिणाम उत्साहजनक हैं क्योंकि वे स्वचालित रूप सेcue phrase (जैसे: हालांकि, बहरहाल, जबकि इत्यादि) सूचना का उपयोग करके अत्याधुनिक सटीकता प्राप्त करते हैं। इन्होंने विधिक पाठ सारांशीकरण में अनुसंधान के लिए एक नया कार्पस तैयार किया है, जिसमें विधिक प्रोक्ति के साथ एनोटेसन के तीन स्तरों- अलंकारिक स्थिति, प्रासंगिकता एवं भाषाई विश्लेषण को प्रस्तुत किया गया है। इस संसाधन की नवीनता और उपयोगिता इस तथ्य में निहित है कि यह पाठसारांशीकरण समुदाय को एक नया सामान्य संसाधन प्रदान करता है जो एकरोचक एवं मूल्यवान डोमेनमें तुलनात्मक अनुसंधान की अनुमतिप्रदान करता है।

5. Sarvanan (2006) ने इस पत्र में, विधिक डोमेन से संबंधित स्वचालित पाठसारांशीकरण कार्य के लिए एक संभाव्य ग्राफिकल मॉडल लागू करने के लिए एक नया विचार प्रस्तावित किया था। इस कार्य में पाठ खनन प्रक्रिया के अंतर्गत् विधिक आलेखों के वाक्यों में उपस्थित अलंकारिक भूमिकाओं की पहचान करना महत्वपूर्ण है। किसी दिए गए विधिक आलेख को सात लेबल वाले घटकों में विभाजितकरने के लिए एक Conditional Random Field (CRF) को लागू किया जाता है और प्रत्येक लेबल उपयुक्त अलंकारिक भूमिकाओं का प्रतिनिधित्व करता है। CRF प्रदर्शन में महत्वपूर्ण सुधार प्रदान करने के लिए अलगअलग विशेषताओं वाले फीचर सेट को नियोजित किया जाता है। इनकी प्रणाली तब अलंकारिक श्रेणियों से संबंधित प्रमुख वाक्यों को निकालने केलिए संरचित डोमेन ज्ञान के साथ एक टर्म डिस्ट्रीब्यूशन मॉडल के अनुप्रयोग से समृद्ध होती है। अंतिम संरचित सारांश को क्षेत्र विशेषज्ञों द्वारा तैयार किए गए आर्दशसारांश के 80% सटीकता स्तर के लगभग देखा गया है।

6. Mehdi (2010) ने इस शोध-पत्र में विधिक पाठ/आलेख सारांशीकरण के लिए सुपरवाइज्ड(पर्यवेक्षित) मशीन लर्निंग दृष्टिकोण को प्रस्तुत किया है। विधिक आलेखों के विश्लेषण एवं सारांशीकरण करने के लिए एवं मशीन लर्निंग प्रयोगों के लिए एक वाणिज्यिक प्रणाली ने इन्हें लगभग 4000 पाठ का एक संग्रह प्रदान किया था। इस संग्रह को उन चुनिंदा स्रोत वाक्यों की पहचान करने के लिए पूर्व-संसाधित किया गया था, जिससे इन्होंने विधिक संरचित डेटा उत्पन्न किया था। अंत में इस संग्रह के साथnaive bayes classifier का उपयोग करके वाक्य वर्गीकरण प्रयोगों का वर्णन किया है।

7. Fillippo (2012) ने पाठ के बारे में विभिन्न प्रकार की सांख्यिकीय सूचनाओं को संयोजित करके नियम बनाने के आधार पर पाठ सारांशीकरण के लिए एक संकर दृष्टिकोण को प्रस्तुत किया है। सुपरवाइज्ड(पर्यवेक्षित) मशीन लर्निंगके विपरीत जहां मानव अंतर्ज्ञान केवल विशेषताओं और एल्गोरिथम चयन परलागू होता है, वहीं मानव अंतर्ज्ञान नियमों की विशेषताओं को व्यवस्थित करने के लिए भी लागू होता

है, जो अभी भी उपलब्ध डेटासेट द्वारा निर्देशित होता है। इन्होंने एक विशेष सारांशीकरण समस्या विधिक डेटा रिपोर्ट के लिए मुहावरें बनाने के लिए अपना दृष्टिकोण प्रस्तुत किया है। साथ ही प्रकरण रिपोर्ट संबंधित मुहावरों प्रकरणों एवं विधिक के लिए इनकमिंग और आउटगोइंग दोनों तरह के उद्धरणों का एक बड़ा डेटासेट तैयार किया है। इन्होंने Ripple down rules के आधार पर एक ज्ञान प्राप्ति ढांचा तैयार किया है और एक समृद्ध नियम भाषा को परिभाषित किया है जिसमें विचाराधीन प्रकरणों के विभिन्न पहलुओं को सम्मिलित किया गया है। इनके द्वारा एक उपकरण विकसित किया गया है, जो वर्तमान प्रकरण के संदर्भ के आधार पर और विभिन्न स्थितियों के लिए अलग-अलग सूचनाओं का उपयोग करके डेटासेट के निरीक्षण और नियमों के निर्माण की सुविधा प्रदान करता है। जिसके द्वारा केवल 23 नियमों के साथ किसी भी स्वचालित विधि की तुलना में काफी उच्च परिशुद्धता(87.4%) प्राप्त कर सकते हैं।

8. Seth (2016) ने इस शोध-पत्र में CaseSummarizer नामक टूल प्रस्तुत किया है जिसमें विधिक आलेखों के औपचारिक पाठ सारांश के अतिरिक्त डोमेन विशिष्ट ज्ञान के साथ संवर्धित शब्द आवृति के आधार पर मानक सारांश विधियों का उपयोग किया जाताहै। तत्पश्चात संक्षिप्ताक्षरों एवं लचीले नियंत्रणों के साथ एक सूचनात्मक इंटरफ़ेस के माध्यम से सारांश प्रदान किए जाते हैं। डोमेनविशेषज्ञों द्वारा प्रदान किए गए सारांश पाठ और प्रतिक्रिया सहित कई अन्य सारांश प्रणालियों के प्रतिकूलROUGE और मानव मूल्यांकन का उपयोग करके इसका मूल्यांकन किया जाता है।

9. Ambedkar (2017) ने बताया है कि विधिक डोमेन में उपलब्ध अधिक मात्रा में ऑनलाइन सूचनाओं ने विधिक पाठ सारांशीकरण को अनुसंधान का एक महत्वपूर्ण क्षेत्र बना दिया है। इस शोधपत्र में उन्होंने विभिन्न पाठ सारांशीकरण तकनीकों का सर्वेक्षण प्रस्तुत किया है। इसमें विधिक पाठ सारांश में चुनौतियों पर विशेष बल दिया गया है क्योंकि यह विधिक क्षेत्र में सबसे महत्वपूर्ण क्षेत्रों में से एक है। पाठ सारांशीकरण में एकल आलेख और बहु-आलेख सारांशीकरण, निष्कर्षण एवं भावात्मक सारांशीकरण तकनीकों का विस्तार से वर्णन किया है। सारांशीकरण में उपयोग किए जाने वाले विभिन्न डेटासेट और मैट्रिक्स पर भी चर्चा की गई है। साथ ही विधिक पाठ सारांशीकरण में उपयोग किए जाने वाले कुछ सॉफ्टवेयर टूल को संक्षेप में वर्णन किया गया है।

10. Varun (2019) द्वारा इस शोध-पत्र में k-mean clustering तकनीक और TF-IDF का उपयोग करके विधिक आलेखों के स्वचालित पाठ सारांशीकरण के लिए एक संकर विधि प्रस्तावित की है। न्यायालय में अपील के लिए अधिवक्ता द्वारा तैयार किए गए केस सारांश के साथ प्रस्तावित विधि से उत्पन्न सारांश की तुलनाROUGE मूल्यांकन मापदंडों का उपयोग करके की गई है।प्रस्तावित प्रणाली द्वारा उत्पन्न सारांश अधिवक्ता द्वारा तैयार किए गए मूल सारांश के समान है और संभवतः आगे भविष्य में संशोधन-सुधार के बाद न्यायालय में इसका उपयोग किया जा सकता है।

11. Laura और Junyi (2019) ने इस शोध-पत्र में एकतरफा अनुबंध जैसे सेवा की शर्तें, आधुनिक डिजिटल जीवन में महत्वपूर्ण भूमिका निभाते हैं। हालांकि, कुछ उपयोगकर्ता उन दस्तावेज़ों(आलेखों) को पढ़ते हैं, जो कि शर्तों को स्वीकार करने से पहले होते हैं, क्योंकि वे बहुत लंबे होते हैं और भाषा बहुत जटिल होती है। तो ऐसे विधिक दस्तावेज़ों को सरल अंग्रेजी में सारांशित करने का कार्य प्रस्तावित किया है, जो उपयोगकर्ताओं को उन शर्तों की बेहतर समझ रखने में सक्षम बनाता है जिन्हें वे स्वीकार कर रहे हैं।

12. Rahul (2020), इस शोध-पत्र के द्वारा टेक्स्ट रैंक एल्गोरिथम का मूल विचार प्रस्तुत किया है जो आलेखों में अलग-अलग वाक्यों की रैंक की गणना करने के लिए पेज रैंक एल्गोरिथम पर आधारित है। इन्होंने वाक्यों को केवल टोकन करके शुरू किया फिर वाक्यों से विभिन्न अवांछित विराम चिन्हों, संख्याओं और अन्य विशेष वर्णों को हटाने का कार्य किया है। तत्पश्चात उन्होंने विराम शब्दोंको हटाकर मूल वाक्यों से साफ और व्यवस्थित वाक्य प्राप्त किए। फिर व्यवस्थित वाक्यों में सभी शब्दों के समतुल्य वेक्टर प्रतिनिधित्व की गणना शुरूकरके पेज रैंक एल्गोरिथम को फीड किया गया, जो सभी वाक्यों को रैंक देता है और शीर्ष रैंक वाले वाक्य आलेख सारांश बनते हैं।

13. Vedant (2021) ने इस शोध-पत्र में भारत के सर्वोच्च न्यायालय द्वारा दिए गए 10,764 निर्णयों के एनोटेट किए गए डेटासेट एवं साथ ही संबंधित हस्तलिखित सारांश(जिसे हेडनोट कहा जाता है) को प्रस्तुत किया है। आलेखों में संक्षिप्ताक्षरों, नाम इकाइयों एवं शब्द विभाजन को वाक्यों में चिंहित कर सामान्य बनाने के लिए पूर्वसंसाधन किया गया है। साथ ही प्रकरणों से जुड़े लोगों और न्यायाधीशों के नाम, निर्णय, निर्णय तिथि, प्रकरण के उद्धरण और निर्णय में संदर्भित वैधानिक कृत्यों जैसी मेटासूचनाओं की भी व्याख्या करते हैं।इसके अलावा इन निर्णयों को स्वचालित रूप से सारांशित करने के लिए एक weakly supervised approach को प्रस्तावित किया गया है। कुछ अच्छे परिणाम प्रस्तावित weakly supervised approach की प्रभावशीलता को दर्शाते हैं जो प्रभावशाली आधारभूत तकनीकों से बेहतर प्रदर्शन करते हैं।

14. Paheli (2021) ने इस शोध-पत्र में DELSumm एल्गोरिथम को प्रस्तावित किया है, जो एक अनसुपरवाइज्ड(गैर-पर्यवेक्षित) एल्गोरिथम है जो विधिक प्रकरण के आलेखों के extractive सारांशीकरण के लिए एक व्यवस्थित रूप से डोमेन ज्ञान को शामिल करता है जिसमें एक अनुकूल सेटअप में विधिक विशेषज्ञों के दिशानिर्देशों को व्यवस्थित रूप से सम्मिलित करने के लिए डिज़ाइन किया गया है। साथ ही इन्होंने उच्चतम न्यायालय प्रकरण आलेखों का विस्तृत प्रयोग किया है। प्रयोगों से पता चलता है कि इनकी प्रस्तावित अनसुपरवाइज्ड एल्गोरिथम ROUGE मैट्रिक्स के संदर्भ में कई प्रभावी आधारभूत विशेषताओं से बेहतर प्रदर्शन करती है जिसमें सामान्य सारांशीकरण एल्गोरिथम और विधिक विशिष्ट दोनों शामिल हैं।

15. Duy-Hung (2021) ने इस शोध-पत्र में, विधिक क्षेत्र में उनके प्रदर्शन को बेहतर बनाने के लिए मौजूदा गहन सारांशीकरण मॉडल को प्रशिक्षित करने के लिए सुदृढ़ीकरण सीखना (Reinforcement Learning) का उपयोग करने संबंधित प्रस्ताव प्रस्तुत किया है। इन्होंने सारांश मॉडल के प्रशिक्षण के लिए अपने दृष्टिकोण से SOTA परिणाम प्राप्त करते हुए PESC डेटासेट के लिए 25.70 % के ROUGE score तक बढ़त बनाई है। उन्होंने अपने दृष्टिकोण को और अधिक मान्य करने के लिए अलग-अलग कॉन्फ़िगरेशन वाले विभिन्न डेटासेट पर बड़े पैमाने पर प्रयोग किये हैं। प्रायोगिक परिणाम बताते हैं कि अतिरिक्त BillSum और विधिक प्रकरण रिपोर्ट डेटासेट पर प्रभावी आधारभूत विशेषताओं की तुलना में यह विधि लगातार बेहतर है।

16. Satyjit (2022) द्वारा भारतीय विधिक पाठ सारांशीकरण के लिए दो अत्याधुनिक स्वतंत्र डोमेन मशीन लर्निंग मॉडल BART एवं PEGASUS के साथ प्रयोग किया गया है। पाठ सामान्यीकरण दृष्टिकोण की प्रभावशीलता को समझने के लिए BART एवं PEGASUS को निष्कर्षण (extractive) और भावात्मक (abstractive) सारांश के संदर्भ में उनके चरणों के माध्यम से रखा जाता है। सारांशित पाठों का मूल्यांकन डोमेन विशेषज्ञों द्वारा कई मापदंडों पर और ROUGE मैट्रिक्स का उपयोग करके किया गया है। यह दर्शाता है कि प्रस्तावित पाठ सामान्यीकरण दृष्टिकोण डोमेन स्वतंत्र मॉडल के साथ विधिक पाठों में प्रभावी है। जिसमें BART मॉडल ने विधिक पाठों को सारांशित करने में अच्छा प्रदर्शन किया है तथा आलेख की लंबाई को 75% तक कम कर दिया है। PEGASUS का उपयोग भावात्मक (abstractive) सारांशीकरण के लिए किया जाता है, लेकिन यह ज्यादातर समय ठीक से काम नहीं करता है। इस प्रकार सामान्यीकरण पद्धति extractive सारांशीकरण के लिए प्रभावी है, लेकिन निश्चित रूप से यह abstractive सारांशीकरण के लिए उतनी उपयोगी नहीं है।

निष्कर्ष –

उपरोक्त सर्वेक्षण से यह ज्ञात होता है कि प्राकृतिक भाषा संसाधन के विधिक पाठ सारांशीकरण के क्षेत्र में भारतीय और विदेशी भाषाओं में शोधकर्ताओं के हर दृष्टिकोण से बहुत सारे अनुसंधान किए जा रहे हैं। प्रस्तुत शोधपत्र में- विधिक पाठ सारांशीकरण प्रणाली के निर्माण से संबंधित पूर्व में किए गए शोधकार्य और उनमें उपयोग की जाने वाली विभिन्न विधियों, तकनीकों और दृष्टिकोणों का संक्षिप्त विवरण प्रस्तुत किया गया है। यह विभिन्न भाषाओं में विधिक डोमेन उपलब्ध होने के कारण विधिक पाठ सारांशीकरण के लिए उपयुक्त और प्रभावी तकनीकों के निर्माण में मदद करेगा।

संदर्भसूची-

- Atefeh Farzindar and Guy Lapalme, 2004, "LetSum, An Automatic Legal Text Summarizing System", Proceeding in T. Gordon(ed.), Legal Knowledge and Information Systems. JURIX 2004: The Seventeenth Annual Conference. Amsterdam: ISO Press, Pp- 11-18.
- Atefeh Farzindar and Guy Lapalme, 2004, "Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles", Association for Computational Linguistics. Barcelona, Spain. In Text Summarization Branches Out, Pp- 27-34.
- Ben Hachey and Claire Grover, March 2005, "Sequence Modelling for sentence classification in a legal summarization system", SAC'05: Proceedings of the 2005 ACM symposium on Applied Computing, Pp-292-296.
- Ben Hachey and Claire Grover, June 2005, "Automatic Legal Text Summarization: Experiments with Summary Structuring", ICAiL'05: Proceedings of the 10th international conference on Artificial Intelligence and Law. Pp- 75-84.
- Ben Hachey and Claire Grover, April 2006, "Extractive Summarization of Legal Texts", Article in Artificial Intelligence and Law. Copyright 2006 Kluwer Academic Publishers, Netherland.

- M. Sarvanan, B. Ravindran and S. Raman, June 2006, "Improving Legal Document Summarization using Graphical Models", Proceedings of the 2006 conference on legal knowledge and information systems. JURIX 2006: The Nineteenth Annual Conference. Pp- 51-60.
- Mehdi Yousfi monod et al, May 2010, "Supervised Machine Learning for Summarizing Legal document", RALI-DIRO rali, Universite de Montrical and NLP Technologies Inc.
- Fillippo Galgani, Paul Compton, and Achim Hoffmann, 2012, "Combining different Summarization Techniques for Legal Text", In proceedings of the workshop on Innovative Hybrid Approaches to the processing of textual data, Avignon, France, ACL. Pp- 115-123.
- Seth Polsley, Pooja Jhunjhunwala and Ruihong Huang, 2016, "CaseSummarizer: A System for Automated Summarization of Legal Texts", In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, Pp- 258-262.
- Ambedkar Kanapala, Rajendra Pamula and Sukomal Pal, March 2019, "Text Summarization from legal documents: A survey", Artificial Intelligence Review, Volume 51, Issue 3, Pp- 371-402. https://doi.org/10.1007/510462-017-9566-2
- Varun Pandya, August 2019, "Automatic Text Summarization of Legal Cases: A Hybrid Approach."
- Laura Manor and Junyi jessy Li, June 2019, "Plain English Summarization of Contracts", Annual Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, MN. ISBN- 978-1-950737-03-1.
- Rahul C Kore et al, May 2020, "Legal Document Summarization using NLP and MI Techniques", International Journal of Engineering and Computer Science, Volume 9, Issue 5, ISSN: 2319-7242.
- Vedant Parikh et al, 2021, "LawSum: A weakly Supervised approach for Indian Legal Document Summarization", arXivi2110.01188v3[cs.CL].
- Paheli Bhattacharya and et al, June 2021, "Incorporating Domain Knowledge for Extractive Summarization of legal case documents", ICAIL'21: Proceedings of the 18th International Conference on Artificial Intelligence and Law, Pp-22-31, https://doi.org/10.1145/3462757.3466092
- Duy-Hung Nguyen et al, December 2021, "Robust Deep Reinforcement Learning for Extractive Legal Summarization", part of the communications in computer and information science book series (CCIS, Volume 1517).
- Satyajit Ghosh, Mousumi Dutta and Tanaya Das, June 2022, "Indian Legal Text Summarization: A Text Normalization based Approach". DOI:10.485501arXiv.2206.06238
- Livio Robaldo et al, April 2019, "Introduction for Artificial Intelligence Processing for legal texts", Artificial Intelligence and Law. @ Springer Nature B.V.2019. https://doi.org/10.1007/s10506-019-09251-2