सतत विकास लक्ष्यों (एसडीजी) में प्रगति को मापना: बजट दस्तावेजों पर नैचुरल लैंग्विज प्रोसेसिंग (एनएलपी) का अनुप्रयोग

रजनी दहिया और शशि कांत द्वारा ^

यह आलेख बजट दस्तावेजों की जांच कर भारत में केंद्र सरकार और चुनिंदा राज्यों में सतत विकास लक्ष्यों (एसडीजी) पर ध्यान केंद्रित करने की प्रगति की जांच करता है। टॉपिक मॉडलिंग का उपयोग करते हुए, मल्टी-लेबल वर्गीकरण के आधार पर प्रत्येक विषय-वस्तु के अंतर्गत एक या अधिक सतत विकास लक्ष्यों को नियत किया जाता है, ताकि यह निर्धारित किया जा सके कि बजट दस्तावेजों में उल्लिखित सतत विकास लक्ष्यों पर नीतियों में प्रणालीबद्ध या संरचनात्मक अंतर हैं या नहीं। परिणाम दर्शाते हैं कि वर्ष 2012 की तुलना में 2023 में केंद्र सरकार और चुनिंदा राज्यों द्वारा सतत विकास लक्ष्यों पर संयुक्त रूप से ध्यान बढ़ा है। यह देखा गया है कि बदलती परिस्थितियों के आधार पर, समय के साथ भिन्न-भिन्न सतत विकास लक्ष्यों पर महत्व में परिवर्तन आता है तथा अलग-अलग सतत विकास लक्ष्यों पर केंद्र सरकार और चुनिंदा राज्यों के बीच उनके फोकस के संदर्भ में काफी विविधता है।

परिचय

सतत विकास लक्ष्य (एसडीजी) 17 महत्वाकांक्षी वैश्विक लक्ष्यों का एक समूह है जो गरीबी, असमानता, जलवायु परिवर्तन और पर्यावरणीय क्षरण जैसी महत्वपूर्ण चुनौतियों से निपटता है। ये लक्ष्य 2012 में रियो+20 सम्मेलन से उभरे और 2015 में संयुक्त राष्ट्र (यूएन) के सभी सदस्य देशों द्वारा औपचारिक रूप से अपनाए गए (अनुबंध में सारणी ए1 इनमें से प्रत्येक लक्ष्य का विस्तृत विवरण प्रदान करती है)।[1] एसडीजी, वर्ष 2000 के सहस्राब्दी विकास लक्ष्यों (एमडीजी) के अगले वाहक हैं, लेकिन वे पिछली सफलताओं और सबक से सीख लेकर आगे बढ़ते हैं। वे वर्ष 2030 तक अधिक धारणीय और निष्पक्ष दुनिया के निर्माण की दिशा में सहयोगी कार्रवाई हेतु एक रूपरेखा प्रदान करते हैं।

एसडीजी के तहत प्रत्येक लक्ष्य के लिए हासिल की गई प्रगति को समाविष्ट करने से यह नीति निर्माताओं, संगठनों और व्यक्तियों को उद्देश्यपूर्ण, परिणामदायक कार्रवाई करने के लिए सशक्त बनाता है। इस दिशा में एक कदम उठाते हुए, एसडीजी इंडिया इंडेक्स का पहला संस्करण दिसंबर 2018 में शुरू किया गया था, जिसमें 13 एसडीजी में 39 लक्ष्यों के 62 संकेतक शामिल किए गए थे।[2] नीति आयोग एसडीजी इंडिया इंडेक्स जारी करता है, जो मंत्रालयों और विभागों द्वारा उपलब्ध कराए गए नवीनतम आंकड़ों का उपयोग करके एसडीजी प्राप्त करने की दिशा में सभी राज्यों और केंद्र शासित प्रदेशों (यूटी) द्वारा की गई प्रगति का आकलन करता है। सूचकांक, सांख्यिकी और कार्यक्रम कार्यान्वयन मंत्रालय (एमओएसपीआई) द्वारा नीति आयोग के परामर्श से विकसित राष्ट्रीय संकेतक रूपरेखा (एनआईएफ) पर आधारित है। प्रत्येक राज्य/ केंद्र शासित प्रदेश के लिए सामान्यीकृत अंकों की गणना की जाती है और किसी लक्ष्य से संबंधित सभी संकेतकों के सामान्यीकृत मूल्यों के अंकगणितीय माध्य का उपयोग करके लक्ष्य के समग्र सूचकांक की गणना की जाती है। प्रत्येक राज्य/ केंद्र शासित प्रदेश के प्रत्येक संकेतक को समान महत्व (मान) दिया जाता है।

यह आलेख एनएलपी में हुई उन्नति का लाभ उठाकर एसडीजी में प्रगति को मापने के लिए एक नया दृष्टिकोण प्रदान करता है। इस दृष्टिकोण का उद्देश्य उभरती बाजार अर्थव्यवस्थाओं में डेटा अंतर को पाटना और बैकवर्ड डेटा उपलब्ध नहीं होने की स्थिति में समय शृंखला का निर्माण करना है (कॉनफोर्टी और अन्य, 2020)। इसके अतिरिक्त, यह मैन्युअल लेबलिंग में लगने वाले अधिक समय और इसके पूर्वाग्रह से ग्रस्त होने की संभावना से परे, विविध स्रोतों से विस्तृत दस्तावेजों की स्वचालित टेक्स्ट लेबलिंग में मदद कर सकता है (गुइसियानो और अन्य, 2022)। इस अध्ययन में कार्यप्रणाली, बजट दस्तावेजों से जानकारी प्राप्त करने के लिए एनएलपी तकनीकों का उपयोग करती है और बजट दस्तावेजों में उल्लिखित एसडीजी पर नीतियों में व्यवस्थित या संरचनात्मक अंतर हैं या नहीं, इसका अध्ययन करने के लिए एक बड़े भाषा मॉडल (एलएलएम) को विनियोजित करती है। डिजिटल

^ लेखक आर्थिक और नीति अनुसंधान विभाग (डीईपीआर) से हैं। हम डॉ. जी.वी. नथनएल को उनकी अंतर्दृष्टिपूर्ण टिप्पणियों के लिए आभार प्रकट करते हैं, जिससे आलेख की गुणवत्ता में काफी सुधार हुआ है। इस आलेख में व्यक्त किए गए विचार लेखकों के हैं और वे आरबीआई के विचारों को नहीं दर्शाते हैं।

[1] संयुक्त राष्ट्र एसडीजी: https://sdgs.un.org/goals.

[2] एसडीजी इंडिया इंडेक्स मेथोडोलॉजी, 2018: https://sdgindiaindex.niti.gov.in/assets/Files/SDG%20India%20Index%202018%20Methodology.docx.

युग में एनएलपी एल्गोरिदम, विशेष रूप से ट्रांसफॉर्मर मॉडल, विस्तृत असंरचित टेक्स्ट डेटा को समझने में सहायक होते हैं, जिससे मशीनों को सूचना को कुशलतापूर्वक संसाधित करने में मदद मिलती है (वासवानी और अन्य, 2017)। एनएलपी की मूल बातें अनुबंध में प्रस्तुत की गई हैं।

भारत में एसडीजी में प्रगति पर उपलब्ध अध्ययनों ने मुख्य रूप से चुनौतियों, कार्यान्वयन और प्रदर्शन की राज्यवार तुलना पर ध्यान केंद्रित किया है (मोहनदास, 2018; पंडा और अन्य, 2018); जबकि कई अध्ययनों ने अन्य देशों के संदर्भों जैसे कि अमेल-ज़ादेह और अन्य, (2021) और मात्सुई और अन्य, (2022) में एसडीजी और एनएलपी के प्रतिच्छेदन का पता लगाया है, यह आलेख बड़े टेक्स्ट डेटासेट के भीतर विषयवस्तु-आधारित एसडीजी वर्गीकरण प्रदान करके साहित्य में योगदान देता है। साथ ही, एसडीजी के मध्य संबंधों का खांका तैयार करने के लिए एनएलपी तकनीकों और नेटवर्क विश्लेषण के मिश्रण का उपयोग करके, यह आलेख उनके बीच अंतर्संबद्धता की संभावना तलाश करता है। यह सह-घटना विश्लेषण महत्वपूर्ण विषयगत संबंधों को प्रकट करता है और एसडीजी के बीच संभावित तालमेल की पहचान करता है (स्मिथ और अन्य, 2021; झोउ और अन्य, 2022)।

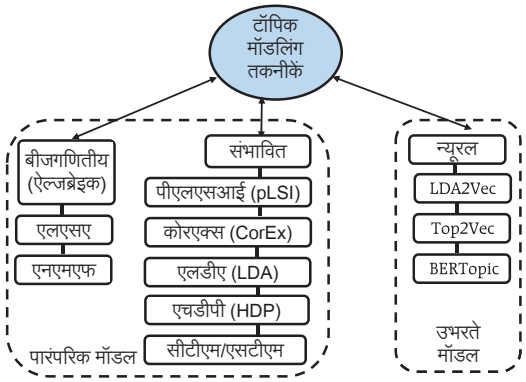
समग्र रूप से, एनएलपी तकनीकें एसडीजी इंडिया इंडेक्स जैसे मात्रात्मक आकलन में योगदान दे सकती हैं। केंद्र सरकार के बजट दस्तावेजों के साथ-साथ उपलब्धता के आधार पर चुनिंदा राज्य सरकारों का विश्लेषण करके, यह आलेख एसडीजी संबंधी सरकार के समग्र वर्णन और संदेश को उजागर करने के लिए एनएलपी का उपयोग करता है। यह विषयगत विश्लेषण के माध्यम से सरकार की प्राथमिकताओं और आख्यानों का गहनतापूर्वक पता लगाता है, बजट दस्तावेजों में आवर्ती विषयों और मुद्दों की पहचान करता है, और महज संख्याओं से परे एक गहन समझ प्रदान करता है। इसमें तुलनात्मक विश्लेषण की भी गुंजाइश है जो यह अंतर्दृष्टि प्रदान करता है कि विभिन्न सरकारी संस्थाएं एसडीजी से कैसे निपटती हैं और समय के साथ सरकार की संचार प्रक्रिया में प्रवृत्तियों का कैसे पता लगाती हैं। तदनुसार, आलेख का शेष भाग निम्नानुसार व्यवस्थित है: खंड II में टॉपिक मॉडलिंग पर चर्चा की गई है; खंड III में डेटा और कार्यप्रणाली पर विस्तार से चर्चा की गई है; खंड IV में परिणामों का विवरण दिया गया है, तथा खंड V में विश्लेषण का समापन किया गया है।

II. एनएलपी और विषय (टॉपिक) मॉडलिंग: कुछ मूल बातें

एनएलपी में विषय मॉडलिंग, पाठ्य दस्तावेजों को उनकी अंतर्निहित अर्थ संरचना के आधार पर व्यवस्थित करने के लिए एक अपर्यवेक्षित शिक्षण दृष्टिकोण है। कलन-गणितीय दृष्टिकोण से, टॉपिक मॉडलिंग तकनीकों को मोटे तौर पर तीन मुख्य श्रेणियों में वर्गीकृत किया जा सकता है: बीजगणितीय, संभावित और न्यूरल मॉडल (वायंस्की और अन्य, 2020; अब्देलराज़ेक और अन्य, 2023)। पहली दो श्रेणियाँ पारंपरिक सांख्यिकी-आधारित विधियों के अंतर्गत आती हैं, जबकि तीसरे एनएलपी में कृत्रिम न्यूरल नेटवर्क को नियोजित करने के अधिक समकालीन दृष्टिकोण को दर्शाया जाता है (चार्ट 1)।

पारंपरिक मॉडलों में, एलडीए दशकों से प्रमुख विकल्प रहा है (ब्लेई और अन्य, 2003)। हालाँकि, कई अध्ययनों ने अपने विषय मॉडलिंग विधि चयन के औचित्य की उपेक्षा करते हुए, बिना किसी विचार-विमर्श के एलडीए को लागू किया है। एलडीए का एक बड़ा दोष यह है कि यह बैग-ऑफ-वर्ड्स (बीओडब्ल्यू) दृष्टिकोण पर निर्भर है, जो शब्दों के बीच अर्थ संबंधों की उपेक्षा करता है (चेन और अन्य, 2023)। सामान्य तौर पर, एलडीए जैसी पारंपरिक विषय मॉडलिंग तकनीकों के लिए जटिल कॉर्पस का प्रसंस्करण-पूर्व, मापदंड का सावधानीपूर्वक चयन (जैसे, विषयों की संख्या निर्धारित करना), उचित मॉडल मूल्यांकन तथा सामान्य बोध और क्षेत्र विशिष्ट का ज्ञान, दोनों पर निर्भर करते हुए उत्पन्न

चार्ट 1: टॉपिक मॉडलों की श्रेणियाँ



टिप्पणी: पारंपरिक मॉडलों में, बीजगणितीय मॉडलों में अंतर्निहित अर्थ विश्लेषण (लेटेंट सिमेंटिक एनालिसिस - एलएसए), और नॉन-नेगेटिव मैट्रिक्स फैक्टराइजेशन (एनएमएफ) शामिल होते हैं। संभावित मॉडलों में प्रॉबबिलिस्टिक लेटेंट सिमेंटिक इंडेक्सिंग (पीएलएसआई), एंकर्ड कोरिलेशन एक्सप्लेनेशन (कोरएक्स) [गैलाघेर और अन्य, 2017], लेटेंट डिरिचलेट एलोकेशन (एलडीए) [ब्ली और अन्य, 2003] के साथ-साथ एलडीए के कई एक्सटेंशन और वेरिएंट शामिल हैं, जिनमें पदानुक्रमित डिरिचलेट प्रक्रिया (एचडीपी), सहसंबद्ध विषय मॉडल (सीटीएम), और संरचनात्मक विषय मॉडल (एसटीएम) शामिल हैं।
स्रोत: चेन, और अन्य, (2023); और लेखकों द्वारा चित्रण।

विषयों की व्याख्या की आवश्यकता होती है। हालिया विकसित न्यूरल मॉडल ने वर्ष 2016 से काफी लोकप्रियता हासिल की है। न्यूरल श्रेणी के उदाहरणों में LDA2Vec (मूडी, 2016), SBM (गेरलाच और अन्य, 2018), डीप एलडीए (भट और अन्य, 2020), Top2Vec (एंजेलोव, 2020) और BERTopic (ग्रूटेंडोर्स्ट, 2022) शामिल हैं। यह प्रवृत्ति हाल के वर्षों में गहन शिक्षण की घातीय प्रगति के साथ संरेखित है।

III. डेटा और कार्यप्रणाली

यह अध्ययन वर्ष 2012 से 2023 तक की अवधि के लिए चुनिंदा भारतीय राज्यों (आंध्र प्रदेश, कर्नाटक, केरल, ओडिशा, तमिलनाडु और पश्चिम बंगाल) और केंद्र सरकार के बजट भाषणों का उपयोग करता है। जैसा कि चार्ट 2 में दर्शाया गया है, संबंधित बजट भाषणों के पाठ के पूर्व-प्रसंस्करण के बाद, अलग-अलग थीम बनाने के लिए BERTopic का इस्तेमाल किया गया क्योंकि इसने एनएमएफ, एलएसए और एलडीए जैसे पारंपरिक मॉडलों से बेहतर प्रदर्शन किया (एगर और अन्य, 2022; चेन और अन्य, 2023)।[3] इन विषय-वस्तुओं को *हगिंग फेस* के प्री-ट्रेन्ड एसडीजी क्लासिफायर मॉडल का उपयोग करके विशिष्ट एसडीजी में वर्गीकृत किया गया था।[4] यह ध्यान रखना महत्वपूर्ण है कि यह अध्ययन पहले 15 एसडीजी पर केंद्रित है, क्योंकि प्री-ट्रेन्ड मॉडल, ओएसडीजी समुदाय से सार्वजनिक रूप से उपलब्ध डेटा के आधार पर परिष्कृत किया गया है (ओएसडीजी, 2022)।

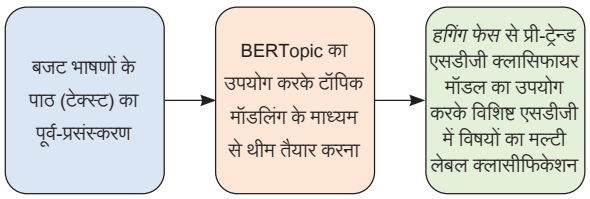
समय के साथ एसडीजी लक्ष्यों पर महत्व देने की यात्रा को मापने के लिए, एनएलपी में एक प्रमुख साधन जिसे टीएफ-आईडीएफ (TF-IDF)[5] के रूप में जाना जाता है, का उपयोग किया जाता है। टीएफ-आईडीएफ के पीछे अंतर्ज्ञान यह है कि किसी शब्द का महत्व दस्तावेजों में उसकी आवृत्ति से प्रतिलोमतः संबंधित है। यह स्कोर दो कारकों को जोड़ता है: शब्द आवृत्ति (टीएफ) जो मापता है कि कोई शब्द किसी दस्तावेज़ में कितनी बार आया है, और प्रतिलोम दस्तावेज़ आवृत्ति (आईडीएफ) जो पूरे कॉर्पस में सामान्य (एक जैसे) शब्दों को ठीक करता है।

टीएफ-आईडीएफ (TF-IDF) स्कोर की गणना, टीएफ (TF) और आईडीएफ (IDF) को गुणा करके की जाती है।

$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D}$$

अनुबंध में टीएफ-आईडीएफ (TF-IDF) के बारे में विस्तृत विवरण दिया गया है।

चार्ट 2: विषय-आधारित एसडीजी वर्गीकरण प्रक्रिया



स्रोत: लेखकों द्वारा चित्रण।

[3] BERTopic बाइडायरेक्शनल एनकोडर रिप्रेजेंटेशंस ट्रांसफार्मर (बीईआरटी) वर्ड एम्बेडिंग के उपयोग के माध्यम से संदर्भ और शब्दार्थ विज्ञान का पता लगाने में प्रभावी है (डेवलिन और अन्य, 2019; अमीन और अन्य, 2022)।

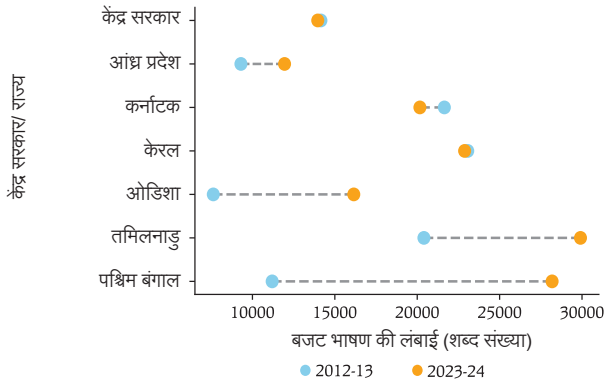
[4] प्री-ट्रेन्ड एसडीजी क्लासिफायर मॉडल यहां उपलब्ध है: https://huggingface.co/jonas/sdg_classifier_osdg.

IV. परिणाम

सबसे पहले, बजट दस्तावेजों में संचार रणनीति के समग्र विकास का विश्लेषण किया गया। शब्द गणना के संदर्भ में यह पाया गया कि केंद्र सरकार, केरल और कर्नाटक के लिए बजट भाषणों की लंबाई में मामूली कमी आई है, जबकि वर्ष 2012-13 की तुलना में 2023-24 में अन्य राज्यों के लिए यह बढ़ गई है (चार्ट 3)।

शब्दों की संख्या से हटकर, बजट में उल्लिखित विषयों की विशिष्ट संख्या का विश्लेषण करने का प्रयास किया गया। विषय

चार्ट 3: बजट भाषण की लंबाई



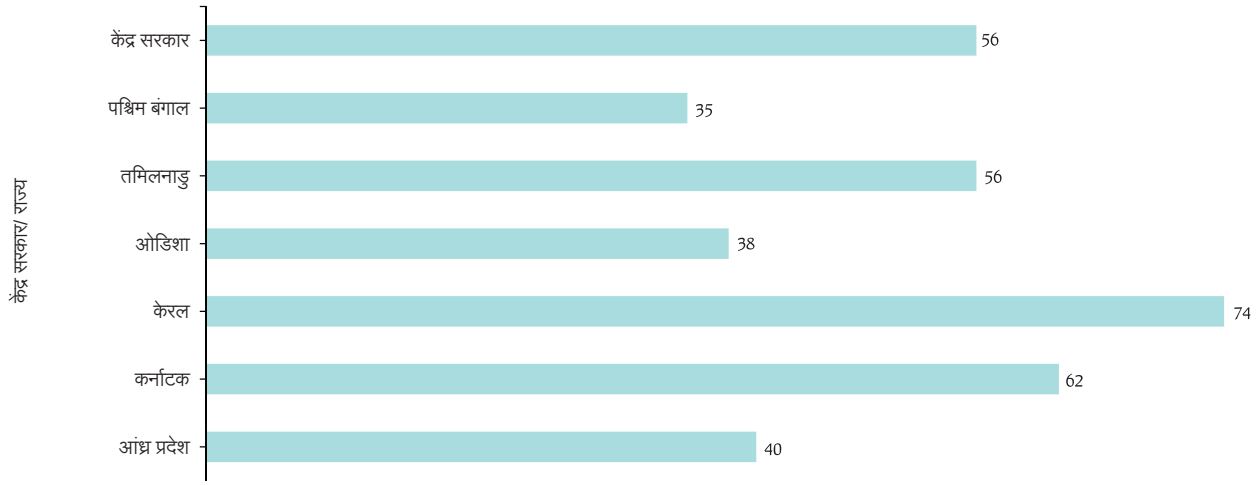
टिप्पणियाँ: 1. वर्ष 2012-13 में कर्नाटक के समान, वर्ष 2023-24 में तमिलनाडु और ओडिशा में कृषि को समर्पित अलग-अलग बजट भाषण थे। विश्लेषण के लिए बजट भाषण एक साथ मिला दिये गए हैं।

2. यह चार्ट डेटा प्री-प्रोसेसिंग से पहले बजट भाषणों की प्रारंभिक शब्द गणना को दर्शाता है।

स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

[5] शब्द आवृत्ति - प्रतिलोम दस्तावेज़ आवृत्ति (टीएफ-आईडीएफ/ TF-IDF)।

चार्ट 4: विषयों की संख्या



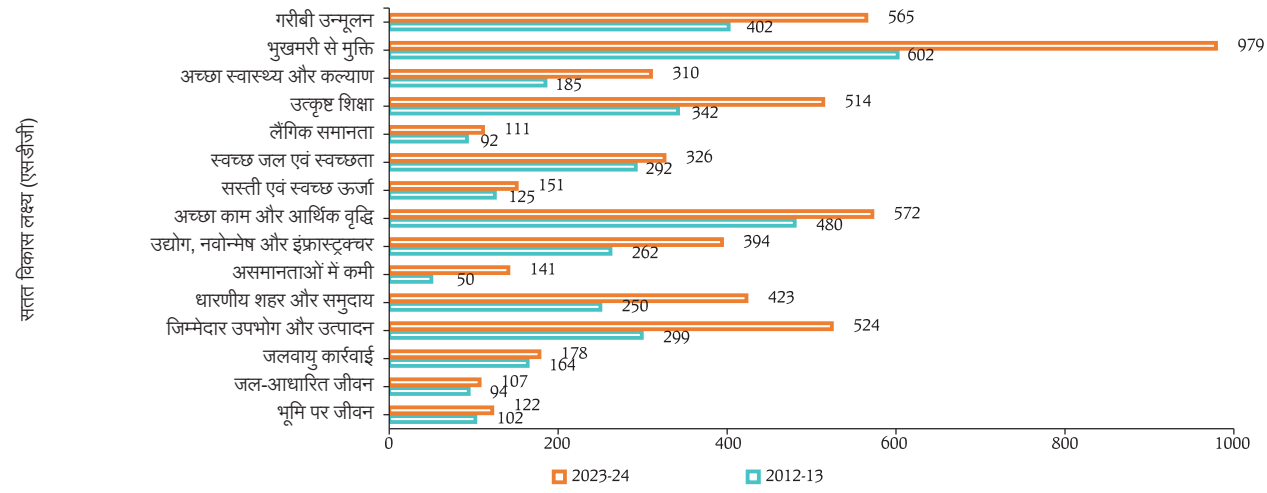
टिप्पणी: थीम (-)1 को हटा दिया गया है क्योंकि इसमें आमतौर पर या तो सामान्य विषय या अप्रासंगिक शब्दों का संयोजन होता है, इसलिए इसे एक आउटलिएर के रूप में माना जाता है (ग्रोटेन्डोर्स्ट, 2022), और टेक्स्ट क्लीनिंग के दौरान कुछ अतिरिक्त विषयों को भी हटा दिया गया था क्योंकि वे एसडीजी के साथ संरेखित नहीं थे। विश्लेषित बजट दस्तावेजों की कुल संख्या 84 थी।
स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

मॉडलिंग के माध्यम से, नमूना राज्यों में केरल के बजट भाषण एसडीजी विषयों में सबसे विविधतापूर्ण रहे, उसके बाद कर्नाटक, केंद्र सरकार और तमिलनाडु का स्थान रहा (चार्ट 4)।

IV.1 समय के साथ सतत विकास लक्ष्यों (एसडीजी) पर महत्व में परिवर्तन

परिणाम दर्शाते हैं कि केंद्र सरकार और चुनिंदा राज्यों द्वारा सभी पंद्रह एसडीजी पर 2012-13 की तुलना में 2023-24 में संयुक्त महत्व बढ़ा है, जो वर्ष 2030 तक इन लक्ष्यों को प्राप्त करने के प्रति उनकी दृढ़ प्रतिबद्धता को दर्शाता है। विशेष रूप से, भुखमरी से मुक्ति (एसडीजी 2), अच्छा स्वास्थ्य और कल्याण (एसडीजी 3), असमानताओं में कमी (एसडीजी 10), धारणीय शहर और समुदाय (एसडीजी 11), और जिम्मेदार उपभोग और उत्पादन (एसडीजी 12) पर महत्व में 2012-13 की तुलना में 2023-24 में उल्लेखनीय वृद्धि देखी गई है (चार्ट 5)।

चार्ट 5: एसडीजी फोकस ट्रांज़िशन (वर्ष 2012-13 की तुलना में 2023-24)



टिप्पणी: सी-टीएफ-आईडीएफ (अनुबंध में विवरण) से प्राप्त आवृत्ति केंद्र सरकार और चुनिंदा राज्यों, दोनों के लिए एकत्रित की जाती है।
स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

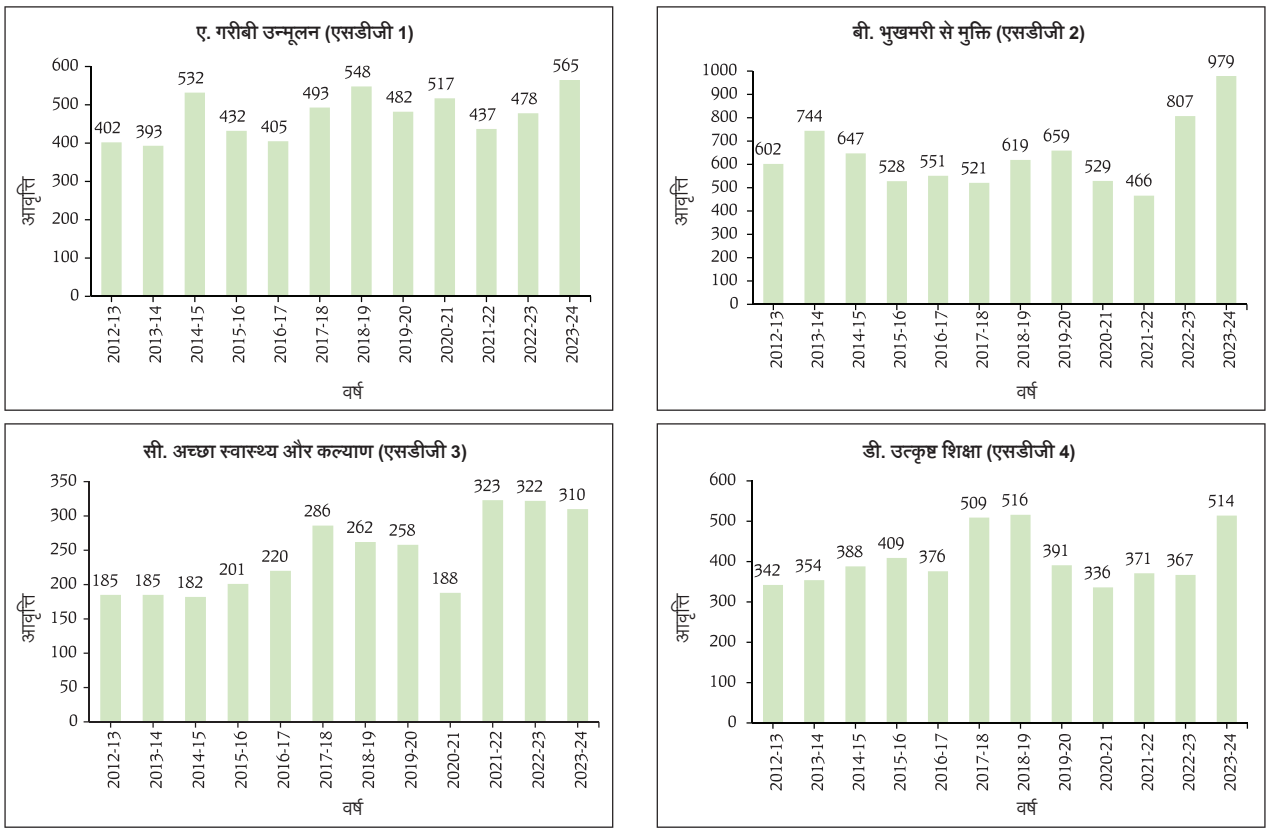
IV.2 सतत विकास लक्ष्यों में रुझान

वर्ष 2012-13 से 2023-24 के दौरान प्रमुख एसडीजी के रुझानों पर गहनतापूर्वक विचार करने के लिए, पहले चार एसडीजी पर केंद्र सरकार और चुनिंदा राज्यों द्वारा संयुक्त महत्व का विश्लेषण किया गया था। इस विश्लेषण से पिछले कुछ वर्षों में फोकस में भिन्नता का पता चलता है। वर्ष 2023-24 में, गरीबी उन्मूलन (एसडीजी 1) और भुखमरी से मुक्ति (एसडीजी 2), दोनों अपने उच्चतम केंद्र-बिंदु पर पहुंच गए (चार्ट 6ए और 6बी)। कोविड-19 महामारी ने अच्छे स्वास्थ्य और कल्याण (एसडीजी 3) पर महत्व को बुरी तरह से प्रभावित किया, इसमें 2021-22 में भारी वृद्धि हुई और यह पिछले नौ वर्षों के अपने उच्चतम स्तर पर पहुंच गया (चार्ट 6सी)। उत्कृष्ट शिक्षा (एसडीजी 4) के लिए, वर्ष 2018-19 और 2023-24, दोनों में शिक्षा क्षेत्र में महत्वपूर्ण सुधार हुए। प्रमुख योजनाएं यथा *समग्र शिक्षा अभियान* [6] (2018) और राष्ट्रीय शिक्षा नीति (2023), इन वर्षों को भारत में शिक्षा के विकास के लिए महत्वपूर्ण मानती हैं (चार्ट 6डी)।

IV.3 सतत विकास लक्ष्यों की अंतर्संबद्धता का निर्धारण: एक सह-घटना विश्लेषण

एसडीजी परस्पर अनन्य नहीं हैं और किसी एसडीजी विशेष पर ध्यान केंद्रित करने से अन्य संबद्ध लक्ष्यों को भी प्राप्त करने में सहायता मिल सकती है। इन सह-घटनाओं का विश्लेषण करने

चार्ट 6: केंद्र सरकार और राज्यों में संयुक्त रूप से प्रमुख सतत विकास लक्ष्यों (एसडीजी) के रुझान



टिप्पणियाँ: 1. यहां प्रयुक्त आवृत्ति मीट्रिक सी-टीएफ-आईडीएफ प्राप्तांक है।

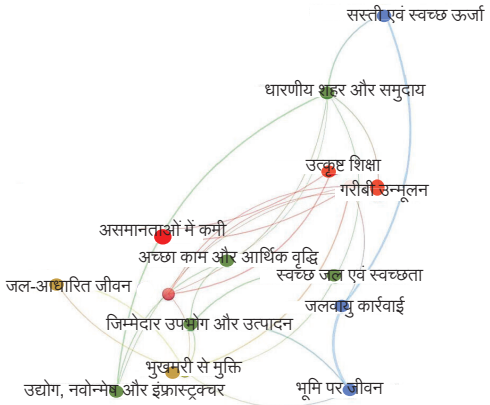
2. यद्यपि बजट भाषण संयुक्त राष्ट्र द्वारा निर्धारित सभी पंद्रह एसडीजी को संबोधित करते हैं, यह शोध पत्र विश्लेषण दृष्टिकोण को प्रदर्शित करने के लिए केवल पहले चार एसडीजी में रुझान दर्शाता है। शेष एसडीजी में रुझानों का पता लगाने के लिए इसी तरह के चित्रण प्रस्तुत किए जा सकते हैं।

3. चुनावी वर्षों में केंद्र सरकार और चुनिंदा राज्यों के लिए, विश्लेषित बजट भाषण अंतिम भाषण होते हैं, अंतरिम भाषण नहीं।

स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

[6] *समग्र शिक्षा योजना* स्कूली शिक्षा के लिए एक एकीकृत योजना है जो शिशु/प्राथमिक कक्षा (प्री-स्कूल) से लेकर कक्षा XII तक के सभी पहलुओं को सम्मिलित करती है। यह योजना स्कूली शिक्षा को एक सतत प्रवाह के रूप में मानती है और सतत विकास लक्ष्य (एसडीजी) 4 के अनुरूप है। कृपया योजना के बारे में https://dsel.education.gov.in/scheme/samagra-shiksha. देखें।

चार्ट 7: अतिव्यापी विषयों के आधार पर एसडीजी की सह-घटनाएँ



टिप्पणी: चार्ट में एसडीजी को जोड़ने वाली रेखाओं की मोटाई उनके बीच साझा किए गए विषयों की संख्या को दर्शाती है, जबकि मोटी रेखाएं अधिक विषयों का प्रतिनिधित्व करती हैं।
स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

से एसडीजी की अंतर्संबद्धता के बारे में बहुमूल्य अंतर्दृष्टि मिलती है, जो एक साथ कई लक्ष्यों को संबोधित करने वाले एकीकृत नीति दृष्टिकोणों की पहचान करने में मदद करती है। एक-समान विषयों के आधार पर एसडीजी की सह-घटना का विश्लेषण चार्ट 7 में प्रस्तुत किया गया है। नवीकरणीय ऊर्जा से संबंधित विषय-वस्तु (थीम) को सस्ती एवं स्वच्छ ऊर्जा (एसडीजी 7), जिम्मेदार उपभोग और उत्पादन (एसडीजी 12), और जलवायु कार्रवाई (एसडीजी 13) के साथ जोड़ा गया था क्योंकि नवीकरणीय ऊर्जा स्वच्छ और सस्ती ऊर्जा प्रदान करने, धारणीय उपभोग प्रथाओं को बढ़ावा देने और जलवायु परिवर्तन के जोखिम को कम करने के लक्ष्य का सीधे समाधान करती है। इसी तरह, वनों और जैव विविधता पर केंद्रित विषय-वस्तु को जलवायु कार्रवाई (एसडीजी 13) और भूमि आधारित जीवन (एसडीजी 15) के साथ जोड़ा गया था। यह भूमि आधारित जीवन को बनाए रखने और जलवायु परिवर्तन के प्रभाव को कम करने में वनों की महत्वपूर्ण भूमिका को दर्शाता है।

IV.4 केंद्र सरकार और राज्यों के बीच नीति संरेखण

अमेल-ज़ादेह और अन्य (2021) ने धारणीयता प्रकटीकरण में पाठ के विश्लेषण के माध्यम से संयुक्त राष्ट्र एसडीजी के साथ संरेखित कंपनियों को चिन्हांकित करने के लिए एनएलपी तकनीकों का उपयोग किया। केंद्र सरकार और राज्यों के विभिन्न एसडीजी के साथ संरेखण को मापने के लिए समान सिद्धांतों को विस्तारित किया जा सकता है। पूर्वानुमान के अनुसार, केंद्र और राज्यों के नीतिगत केंद्र-बिंदु के अलग-अलग होने की उम्मीद है और समय के साथ भिन्न भी हो सकता है। अभिशासन के विषयों को संविधान तीन भागों में विभाजित करता है, यानी संघ, राज्य और समवर्ती सूचियाँ। यह राष्ट्रीय और उप-राष्ट्रीय सरकारों को समवर्ती सूची में कुछ सीमा तक अतिव्यापन के साथ अपने क्षेत्राधिकार में विषयों पर कार्य करने की स्वतंत्रता और वातावरण देता है जहां दोनों एक-दूसरे की नीतियों में सहयोग कर सकते हैं और पूरक बन सकते हैं। राज्य अपनी स्थानीय परिस्थितियों के अनुरूप नीतियों को परिष्कृत कर सकते हैं। इस परिकल्पना की पुष्टि तब होती है, जब यह अध्ययन दो प्रमुख घटकों का उपयोग करके 15 एसडीजी की आयामीता को घटाकर 2-आयामी प्लेन तक लाता है। प्रत्येक बजट दस्तावेज़ को विभिन्न एसडीजी पर इसके महत्व के संदर्भ में 15x1 वेक्टर द्वारा दर्शाया जा सकता है। नीतियों के रुख की कल्पना करने के लिए, पहले दो प्रधान घटकों को प्लॉट करके आयामीता को 15-डी स्पेस से 2-डी स्पेस में घटाया जाता है, जो कुल भिन्नता के 47 प्रतिशत हिस्से का प्रतिनिधित्व करते हैं (सारणी 1)।

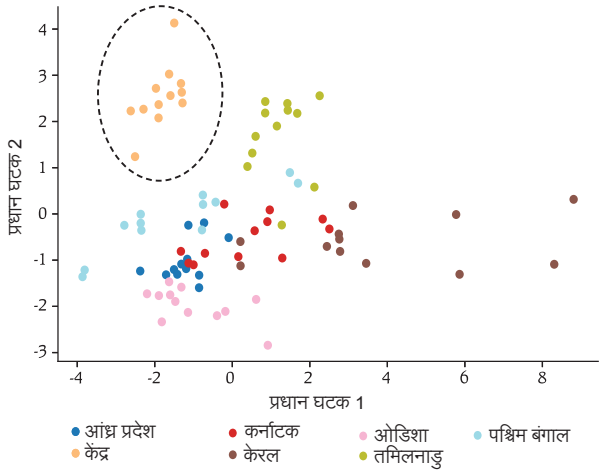
दृश्य चित्रण दर्शाता है कि केंद्र सरकार (काले बिंदुओं द्वारा चिन्हांकित समूह) और राज्यों के बजट दस्तावेजों में शामिल विषयों को एक-दूसरे से अलग किया गया है (चार्ट 8)। जबकि कुछ राज्यों की नीतियों में पिछले कुछ वर्षों में कम भिन्नता दिखाई देती है, वहीं अन्य राज्यों में एसडीजी पर महत्व देने में काफी

सारणी 1: क्रमिक प्रधान घटक (पीसी) द्वारा वर्णित संचयी घट-बढ़

पीसी 1	पीसी 2	पीसी 3	पीसी 4	पीसी 5	पीसी 6	पीसी 7	पीसी 8	पीसी 9	पीसी 10
0.32	0.47	0.60	0.67	0.74	0.80	0.84	0.88	0.91	0.93
पीसी 11	**पीसी 12**	**पीसी 13**	**पीसी 14**	**पीसी 15**					
0.96	0.97	0.98	0.99	1.00					

टिप्पणी: मान दो दशमलव बिंदुओं तक पूर्णांकित किए गए हैं।
स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

चार्ट 8: प्रधान घटक विश्लेषण



टिप्पणी: प्रत्येक राज्य/केंद्र सरकार को 12 बिंदुओं द्वारा दर्शाया गया है, जो 2012-13 से 2023-24 तक के वर्षों को इंगित करते हैं।

स्रोत: केंद्र सरकार और राज्य बजट भाषण - वर्ष 2012-13 से 2023-24; और लेखकों की गणना।

विविधता दिखाई देती है। राज्यों की तुलना में केंद्र सरकार ने पिछले कुछ वर्षों में सबसे कम भिन्नता दर्शायी है (सारणी 2)। ये अंतर बताते हैं कि राज्य स्तर की नीतियों को उनकी स्थानीय स्थितियों को ध्यान में रखते हुए तैयार किया जा सकता है।

V. निष्कर्ष

डेटा विश्लेषण को बढ़ाने के लिए एनएलपी एक अमूल्य साधन बन गया है और इसे नीति निर्माण में तेजी से लागू किया जा रहा है। यह आलेख भारतीय संदर्भ में सतत विकास लक्ष्यों (एसडीजी) के लिए प्रगति को मापने हेतु एनएलपी के एक नव उपयोग का प्रस्ताव रखता है, जो मौजूदा मात्रात्मक दृष्टिकोणों के लिए सहायक है। केंद्र सरकार और चुनिंदा राज्यों के लिए किए गए विश्लेषण से संकेत मिलता है कि एसडीजी पर उनका संयुक्त ध्यान वर्ष 2012-13 की तुलना में 2023-24 में अधिक था, जो 2030 तक इन लक्ष्यों को प्राप्त करने के लिए एक मजबूत प्रतिबद्धता की ओर इंगित करता है। समय के साथ बदलती परिस्थितियों जैसे कि कोविड-19 महामारी के दौरान स्वास्थ्य पर अधिक ध्यान देने के आधार पर विभिन्न एसडीजी पर महत्व अलग-अलग रहा है। यह भी पाया गया है कि एसडीजी परस्पर अनन्य नहीं हैं और उनके मध्य महत्वपूर्ण अंतर्संबद्धता प्रदर्शित होती है। इसलिए, किसी एसडीजी विशेष को लक्षित करने से अन्य संबद्ध लक्ष्यों को प्राप्त करने में भी मदद मिल सकती है। इसके अलावा, यह दृष्टिकोण केंद्र सरकार और चुनिंदा राज्यों के बीच एसडीजी

सारणी 2: बजट दस्तावेजों में एसडीजी प्रतिनिधित्व में क्लस्टर के आकार

क्षेत्र	औसत समूह (क्लस्टर) आकार
केरल	3.50
कर्नाटक	2.84
तमिलनाडु	2.33
पश्चिम बंगाल	2.26
आंध्र प्रदेश	2.06
ओडिशा	1.93
केंद्र सरकार	1.78

टिप्पणी: चार्ट 8 में औसत क्लस्टर आकार की गणना, प्रत्येक क्लस्टर के केंद्रक से प्रत्येक डेटा बिंदु की यूक्लिडियन दूरी का उपयोग करके की जाती है।

स्रोत: लेखकों की गणना।

पर ध्यान केंद्रित करने में विविधता की जांच करने में सहायता करता है। पारंपरिक मात्रात्मक माप से आगे बढ़कर, इस आलेख में प्रस्तुत सतत विकास लक्ष्यों पर सापेक्ष महत्व का विश्लेषण - समय के साथ विकास, अंतर्संबद्धता और सापेक्ष महत्व के बारे में महत्वपूर्ण नीतिगत अंतर्दृष्टि प्रदान करता है।

संदर्भ

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2023). Topic Modeling Algorithms and Applications: A Survey. *Information Systems, 112, 102131.*

Amel-Zadeh, A., Chen, M., Mussalli, G., and Weinberg, M. (2021). NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals. *Social Science Research Network (SSRN).*

Amin, A., Hassan, S., Alm, C., and Huenerfauth, M. (2022). Using BERT Embeddings to Model Word Importance in Conversational Transcripts for Deaf and Hard of Hearing Users. *ResearchGate Publication.* https://doi.org/10.13140/RG.2.2.28272.33289

Anastasopoulos, L., Moldogazi, T., and Scott, T. (2017). Computational Text Analysis for Public Management Research. *SSRN Electronic Journal, doi:10.2139/ssrn.3269520.*

Angelov, D. (2020). Top2vec: Distributed Representations of Topics. *arXiv 2020, arXiv:2008.09470.*

Antonellis, I., and Gallopoulos, E. (2006). Exploring Term-Document Matrices from Matrix Models in Text Mining. arXiv: cs/0602076.

Bhat, M. R., Kundroo, M. A., Tarray, T. A., and Agarwal, B. (2020). Deep LDA: A New Way to Topic Model. *Journal of Information Optimization Sciences, 41, 823–834.*

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3, 993–1022.*

Camacho-Collados, J., and Pilehvar, M. T. (2017). On the Role of Text Pre-processing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *arXiv, 1707.01780.*

Cavnar, W. B., and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. *Proceedings of SDAIR-94, 161–175.*

Chen, W., Rabhi, F., Liao, W., and Al-Qudah, I. (2023). Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. *Electronics, 12(12), 2605.*

Christian, H., Agus, M., and Suhartono, D. (2016). Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications, 7, 285.*

Conforti, C., Hirmer, S., Morgan, D., Basaldella, M., and Ben Or, Y. (2020). Natural Language Processing for Achieving Sustainable Development: The Case of Neural Labelling to Enhance Community Profiling. *ACL Anthology.*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics.*

Egger, R., and Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology, 7.*

Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. (2017). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Transactions of the Association for Computational Linguistics, 5, 529–542.*

Gerlach, M., Peixoto, T. P., and Altmann, E. G. (2018). A Network Approach to Topic Models. *Science Advances, 4, eaaq1360.*

Grootendorst, M. (2022). BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *arXiv 2022, arXiv:2203.05794.*

Guisiano, J. E., Chiky, R., and De Mello, J. (2022). SDG-Meter: A Deep Learning Based Tool for Automatic Text Classification of the Sustainable Development Goals. *United Nations Environment Program, Paris, France.*

Hachaj, T., and Ogiela, M. R. (2018). What Can Be Learned from Bigrams Analysis of Messages in Social Network? In *2018 11th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)* [pp. 1-4]. Beijing, China. doi:10.1109/CISP-BMEI.2018.8633108.

Khyani, D., and B S, S. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology, 22, 350-357.*

Kurniasih, A., and Manik, L. P. (2022). On the Role of Text Pre-processing in BERT Embedding-based DNNs for Classifying Informal Texts. *International Journal of Advanced Computer Science and Applications (IJACSA), 13(6).*

Matsui, T., Suzuki, K., Ando, K., Kitai, Y., Haga, C., Masuhara, N., and Kawakubo, S. (2022). A Natural Language Processing Model for Supporting Sustainable Development Goals: Translating Semantics, Visualizing Nexus, and Connecting Stakeholders. *Sustain Sci, 17, 969–985.*

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to make Lda2vec. *arXiv, arXiv:1605.02019.*

Mohandas, P. (2018). Sustainable Development Goals (SDGs) - Challenges for India. *Indian Journal of Public Health Research & Development, 9*(1), 1. https://doi.org/10.5958/0976-5506.2018.00172.9

OSDG, UNDP IICPSD SDG AI Lab, and PPMI (2022). OSDG Community Dataset. *Zenodo.* https://doi.org/10.5281/zenodo.6831287.

Panda, R., Sethi, M., and Agrawal, S. (2018). Sustainable Development Goals and India: A Cross-Sectional Analysis. *OIDA International Journal of Sustainable Development, 11*(11), 79-90. Retrieved from https://ssrn.com/abstract=3308074

Sharma, D., and Jain, S. (2015). Evaluation of Stemming and Stop Word Techniques on Text Classification Problem. *International Journal of Scientific Research in Computer Science and Engineering, 3*(2), Page Range. ISSN: 2320-7639. Retrieved from www.isroset.org.

Smith, T. B., Vacca, R., Mantegazza, L., *et al.* (2021). Natural Language Processing and Network Analysis provide Novel Insights on Policy and Scientific Discourse around Sustainable Development Goals. *Scientific Reports, 11*(1), 22427. https://doi.org/10.1038/s41598-021-01801-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. *In Proceedings of the Conference on Neural Information Processing Systems.*

Vayansky, I., and Kumar, S. A. P. (2020). A Review of Topic Modeling Methods. *Information Systems, 94, 101582.*

Webster, J., and Kit, C. (1992). Tokenization as the Initial Phase in NLP. *Proceedings of the 14th Conference on Computational linguistics - Volume 4,* 1106-1110. https://doi.org/10.3115/992424.992434.

Zhou, X., Jain, K., Moinuddin, M., and McSharry, P. (2022). Using Natural Language Processing for Automating the Identification of Climate Action Interlinkages within the Sustainable Development Goals. *In Association for the Advancement of Artificial Intelligence (AAAI) 2022 Fall Symposium on the Role of AI in Responding to Climate Challenges.*

अनुबंध

सारणी ए1: सतत विकास लक्ष्यों (एसडीजी) का विवरण

सतत विकास लक्ष्य (एसडीजी)	विवरण
1. गरीबी उन्मूलन	हर जगह सभी रूपों में गरीबी को समाप्त करें।
2. भुखमरी से मुक्ति	भुखमरी की समाप्ति, खाद्य सुरक्षा और धारणीय कृषि को बढ़ावा
3. अच्छा स्वास्थ्य और कल्याण	प्रत्येक आयुवर्ग के लोगों के लिए स्वस्थ जीवन सुनिश्चित करना तथा कल्याण को बढ़ावा देना।
4. उत्कृष्ट शिक्षा	समावेशी एवं समतापूर्ण उत्कृष्ट शिक्षा सुनिश्चित करना तथा आजीवन सीखने के अवसरों को बढ़ावा देना।
5. लैंगिक समानता	लैंगिक समानता प्राप्त करना और सभी महिलाओं और लड़कियों को सशक्त बनाने के लिए प्रयास करना।
6. स्वच्छ जल एवं स्वच्छता	सभी के लिए जल एवं स्वच्छता की उपलब्धता एवं स्थायी प्रबंधन सुनिश्चित करना।
7. सस्ती एवं स्वच्छ ऊर्जा	सभी के लिए सस्ती, विश्वसनीय, धारणीय और आधुनिक ऊर्जा तक पहुंच सुनिश्चित करना।
8. अच्छा काम और आर्थिक वृद्धि	सभी के लिए सतत, समावेशी और धारणीय आर्थिक वृद्धि, पूर्ण और उत्पादक रोजगार तथा समुचित कार्य को बढ़ावा देना।
9. उद्योग, नवोन्मेष और इंफ्रास्ट्रक्चर	मजबूत इंफ्रास्ट्रक्चर का निर्माण, समावेशी और धारणीय औद्योगिकीकरण को प्रोत्साहित करना और नवोन्मेष को बढ़ावा देना।
10. असमानताओं में कमी	देशों के भीतर और देशों के बीच असमानता कम करना।
11. धारणीय शहर और समुदाय	शहरों और मानव बस्तियों को समावेशी, सुरक्षित, समुत्थानशील और धारणीय बनाना।
12. जिम्मेदार उपभोग और उत्पादन	धारणीय उपभोग और उत्पादन स्वरूप सुनिश्चित करना
13. जलवायु कार्रवाई	जलवायु परिवर्तन और उसके प्रभावों से निपटने के लिए तत्काल कार्रवाई करना
14. जल-आधारित जीवन	सतत विकास के लिए महासागरों, समुद्रों और समुद्री संसाधनों का संरक्षण और सतत उपयोग करना।
15. भूमि पर जीवन	स्थलीय पारिस्थितिकी तंत्रों के सतत उपयोग को संरक्षित, पुनर्स्थापित और बढ़ावा देना, वनों का सतत प्रबंधन करना, मरुस्थलीकरण से निपटना, भूमि क्षरण को रोकना और उसका बचाव तथा जैव विविधता की हानि को रोकना।
16. शांति, न्याय और मजबूत संस्थाएँ	सतत विकास के लिए शांतिपूर्ण और समावेशी समाज को बढ़ावा देना, सभी के लिए न्याय तक पहुंच सुनिश्चित करना, तथा सभी स्तरों पर प्रभावी, जवाबदेह और समावेशी संस्थाओं का निर्माण करना।
17. लक्ष्यों के लिए साझेदारी	कार्यान्वयन के साधनों को मजबूत करना और सतत विकास के लिए वैश्विक साझेदारी को पुनर्जीवित करना।

स्रोत: https://sdgs.un.org/goals.

I. एनएलपी पर परिचयात्मक लेख

क. शब्दावली

एनएलपी में विश्लेषण की मूल इकाई "टर्म" है, जो या तो "विकास" जैसा एक शब्द हो सकता है या "एन-ग्राम (n-grams)" के रूप में जाना जाने वाला शब्दों का एक क्रम हो सकता है (कैवनार और अन्य, 1994)। जबकि शब्द अपने मूल अर्थ को बनाए रखते हैं, n-grams, विश्लेषण के लिए एकल इकाई के रूप में शब्दों के समूहों का प्रतिनिधित्व करते हैं। इसके उदाहरणों में "अच्छा स्वास्थ्य" और "उत्कृष्ट शिक्षा" शामिल हैं जो "bi-grams" हैं (हचज और अन्य, 2018) जिन्हें अक्सर एनएलपी कार्यों के अंतर्गत एकल शब्दों के रूप में विश्लेषित किया जाता है।

दस्तावेज़, वाक्यों एवं अनुच्छेदों से लेकर संपूर्ण साहित्यिक कृतियों तक कुछ भी हो सकते हैं और आम तौर पर पाठ का विश्लेषण करते समय विश्लेषण की प्राथमिक इकाई होते हैं। यहाँ वित्तीय वर्ष 2012 से 2023 के लिए केंद्र सरकार और छह राज्यों के बजट भाषणों (दस्तावेज़ों) का विश्लेषण किया गया है।

अंततः, एक कॉर्पस विश्लेषण किए गए दस्तावेजों का एक संग्रह है, जो 'डेटा सेट' के समतुल्य है। इस मामले में, कॉर्पस बजट भाषणों का समूह है। कॉर्पस में कई दस्तावेज़ शामिल हैं, जिनमें से प्रत्येक में शब्द सम्मिलित हैं। अनास्तासोपोलोस और अन्य, (2017) ने गणितीय संकेतन में एनएलपी प्रसंस्करण पदानुक्रम को बड़े सुव्यवस्थित ढंग से सारांशित किया (चार्ट ए1)।

चार्ट ए1: एनएलपी प्रसंस्करण पदानुक्रम

शब्द (टर्म) ⊆ दस्तावेज़ ⊆ कॉर्पस

स्रोत: अनास्तासोपोलोस और अन्य, (2017); और लेखकों द्वारा चित्रण।

बी. टेक्स्ट से डेटा तक

एनएलपी में टेक्स्ट प्री-प्रोसेसिंग एक महत्वपूर्ण चरण है (कुर्नियासिह और अन्य, 2022)। इसमें प्रारंभिक टेक्स्ट डेटा को परिष्कृत कर संख्याओं में बदलने के लिए कई महत्वपूर्ण चरण शामिल हैं। प्रारंभिक चरण टोकनाइजेशन है जहां टेक्स्ट को अलग-अलग शब्दों या टोकन में विभाजित किया जाता है (वेबस्टर और अन्य, 1992)। सभी वर्णों को लोअरकेस में

परिवर्तित करके टेक्स्ट की स्थिरता सुनिश्चित करने के लिए लोअरकेसिंग आवश्यक है, जिससे विश्लेषण में "budget" और "Budget" समान हो जाते हैं। स्टॉप वर्ड रिमूवल (शर्मा और अन्य, 2015) "द/दि", "इज़", "आर", आदि जैसे सामान्य, कुछ खास जानकारी नहीं देने वाले शब्दों को अलग करता है, ताकि अर्थ व्यक्त करने वाले शब्दों पर ध्यान केंद्रित रहे। विराम चिह्नों, जैसे कि पूर्ण विराम या अल्पविराम को हटाने से टेक्स्ट की पठनीयता में सुधार होता है और अनावश्यक अव्यवस्था दूर होती है। अंत में, स्टेमिंग या लेमेटाइजेशन (ख्यानी और अन्य, 2021) शब्दों को उनके मूल रूपों में सरलीकृत करता है, और यह सुनिश्चित करता है कि समान अर्थ वाले शब्दों को उसी तरह से दर्शाया जाए; इस प्रकार "budgeting", "budgeted" और "budgets" सभी "बजट (budget)" बन जाते हैं (चार्ट ए2)।

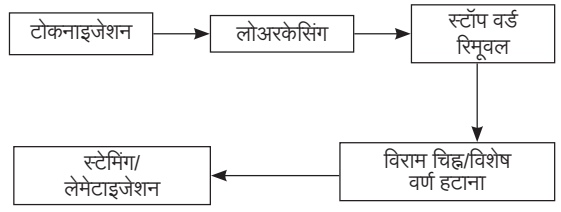
सी. डॉक्यूमेंट टर्म मैट्रिक्स (डीटीएम)

एनएलपी और टेक्स्ट विश्लेषण के लिए डीटीएम एक महत्वपूर्ण साधन है, जो विश्लेषण के लिए दस्तावेजों को एक संरचित संख्यात्मक प्रारूप में परिवर्तित करता है (एंटोनेलिस और अन्य, 2006; और अनास्तासोपोलोस और अन्य, 2017)। यह प्रत्येक दस्तावेज़ को पंक्तियों (रो) में और विशिष्ट शब्दों को स्तंभों (कॉलम) में दर्शाता है, दस्तावेज़ों में शब्द आवृत्ति या उपस्थिति को इंगित करने के लिए मान का उपयोग करता है। खानों (सेल्स) को भरने के लिए विभिन्न तरीकों का उपयोग किया जा सकता है, जिनमें टीएफ (शब्द आवृत्ति), बाइनरी (उपस्थिति/अनुपस्थिति), और टीएफ-आईडीएफ (शब्द आवृत्ति को शब्द विशिष्टता के साथ जोड़ना) शामिल हैं [क्रिश्चियन और अन्य, 2016]। इनमें से सबसे आम टीएफ-आईडीएफ है जिस पर अगले खंड में विस्तार से चर्चा की जाएगी।

डी. टीएफ-आईडीएफ का संक्षिप्त विवरण

टीएफ-आईडीएफ के पीछे मुख्य अंतर्ज्ञान यह है कि किसी शब्द का महत्व दस्तावेज़ों में उसकी आवृत्ति से प्रतिलोमतः

चार्ट ए2: टेक्स्ट की प्रसंस्करण पूर्व प्रक्रिया



स्रोत: कैमाचो-कोलाडोस, जे. और अन्य, (2017); और लेखकों द्वारा चित्रण।

संबंधित है। यह स्कोर दो कारकों को जोड़ता है: शब्द आवृत्ति (टीएफ) मापता है कि दस्तावेज़ में कोई शब्द कितनी बार प्रयुक्त होता है, जबकि प्रतिलोम दस्तावेज़ आवृत्ति (आईडीएफ) पूरे कॉर्पस में सामान्य शब्दों को ठीक करता है। आईडीएफ की गणना इस प्रकार की जाती है:

इस संदर्भ में, t उस शब्द को दर्शाता है जिसके लिए यह शोध पत्र समानता का आकलन करने का लक्ष्य रखता है, और N का अर्थ है कॉर्पस (D) में कुल दस्तावेजों की संख्या (d)। हर/विभाजक (डिनामनेटर), उन दस्तावेजों की संख्या से मेल खाता है जहां शब्द t मौजूद है।

जब कॉर्पस में शब्द न हों, तो शून्य-से-विभाजन वाली त्रुटियों की रोकथाम के लिए, आईडीएफ गणनाएं आम तौर पर शब्द वाले दस्तावेज़ों की संख्या में 1 जोड़ती हैं, जिससे हर/विभाजक को प्रभावी रूप से (1 + गिनती) में समायोजित किया जाता है। लोकप्रिय लाइब्रेरी scikit-learn इस सूत्र को निम्नानुसार संशोधित करके हल करती है:

$$IDF_{t,D} = log\left(\frac{N}{count(d \epsilon D: t \epsilon d)}\right) \qquad \text{Equation (1)}$$

इस प्रकार, आईडीएफ के साथ कभी-कभार प्रयुक्त होने वाले शब्द उभर कर सामने आते हैं, तथा दस्तावेजों के भीतर अप्रत्यक्ष अंतर्दृष्टि का पता चलता है।

$$IDF_t = 1 + log\frac{(1+n)}{(1+DF_t)} \qquad \textit{Equation (2)}$$

कड़ियों को जोड़ना: टीएफ-आईडीएफ

टीएफ और आईडीएफ को गुणा करके टीएफ-आईडीएफ प्राप्तांक की गणना की जाती है।

$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D} \qquad \text{Equation (3)}$$

किसी दस्तावेज़ में किसी शब्द की प्रासंगिकता उसके टीएफ-आईडीएफ प्राप्तांक से परिलक्षित होती है, जिसमें 0 न्यूनतम महत्व को दर्शाता है और उच्च प्राप्तांक बढ़ते महत्व को दर्शाते हैं। यह आलेख सी-टीएफ-आईडीएफ का उपयोग करता है जो एक श्रेणी-आधारित टीएफ-आईडीएफ प्रक्रिया है जिसका उपयोग पाठ्य दस्तावेज़ों से उस श्रेणी के आधार पर विशेषताएँ उत्पन्न करने के लिए किया जा सकता है जिसमें वे हैं [7]।

[7] https://github.com/MaartenGr/cTFIDF