*शब्दों के सामर्थ्य से मुद्रास्फीति का स्पष्ट संकेत देना**

मीडिया, जो सूचना के प्रसार का एक महत्वपूर्ण माध्यम है, जनता के मनोभाव एवं अपेक्षाओं को प्रभावित करने की क्षमता रखता है। इस लेख में भारत में खुदरा मुद्रास्फीति के विशिष्ट संदर्भ में ऑनलाइन प्रिंट मीडिया से उठाई गई अत्यधिक दोहराई जाने वाली अव्यवस्थित सूचनाओं का उपयोग किया गया है। मनोभाव वर्गीकरण के लिए व्यापक रूप से उपयोग की जानेवाली तकनीक, सपोर्ट वेक्टर मशीन (एसवीएम) क्लासिफायर का उपयोग करके समाचारों से मनोभाव निकाला जाता है और इस प्रकार मनोभाव सूचकांक तैयार किया जाता है। प्रयोगसिद्ध परिणाम के अनुसार मीडिया का मनोभाव सूचकांक, मुद्रास्फीति को भली-भांति दर्शाता है। इसकी दिशा की सटीकता उच्च एवं आंकड़ों की दृष्टि से महत्वपूर्ण है। इसके अलावा, ग्रेंजर कारण-कार्य संबंध परीक्षण के नतीजे भी दर्शाते हैं कि मनोभाव सूचकांक में खुदरा मुद्रास्फीति का अनुमान लगाने की पर्याप्त क्षमता है।

परिचय

मीडिया, जो सूचना के प्रचार का एक महत्वपूर्ण माध्यम है, जनता के मनोभाव एवं अपेक्षाओं को प्रभावित करने की क्षमता रखता है, जिसका समष्टि-आर्थिक नतीजों के साथ महत्वपूर्ण अंतर-संबंध हो सकता है। बिग डेटा तकनीकों का उपयोग करके, दुनियाभर में शोधकर्ताओं द्वारा इस तरह के नए डेटा में निहित सूचना का फायदा उठाया जा रहा है, ताकि एक समुचित वैकल्पिक संकेतक तैयार किया जा सके, जिससे समष्टि-आर्थिक चरों का अनुमान लगाने में आसानी हो सके। यह लेख भारत में खुदरा मुद्रास्फीति के विशिष्ट संदर्भ में मीडिया के मनोभाव का विश्लेषण करने के जरिए इस प्रगतिशील साहित्य में योगदान करता है।

अत्यधिक दोहराई जाने वाली एवं बहुतायत वाली अव्यवस्थित सूचना का विश्लेषण करने के लिए बिग डेटा टूल्स का प्रयोग आवश्यक है, जैसे मशीन लर्निंग (एमएल) और नैचुरल लैंग्वेज प्रॉसेसिंग (एनएलपी) तकनीक। एसवीएम क्लासिफायर का उपयोग करके समाचारों से मनोभाव निकाला जाता है, जो मनोभाव वर्गीकरण के लिए व्यापक रूप से उपयोग किया गया तकनीक है। भारत में खुदरा मुद्रास्फीति के विशिष्ट संदर्भ में, हम ऑनलाइन प्रिंट मीडिया समाचारों एवं रिपोर्टों से उठाई गई अव्यवस्थित समाचार विषय-वस्तु का उपयोग करते हैं। इस तरह की अत्यधिक दोहराई जाने वाली सूचना (दैनिक समाचार) की उपलब्धता की बदौलत लगभग तत्काल आधार पर सूचकांक विकसित किया जा सकेगा।

इस पृष्ठभूमि में, (i) मुद्रास्फीति के बारे में समाचारों में निहित मनोभाव को निकालने, (ii) मनोभावों को एकत्रित करने एवं मनोभाव सूचकांक विकसित करने, तथा (iii) मनोभाव और मुद्रास्फीति के बीच के संबंध को अनुभव के आधार पर परखने का प्रयास किया गया है।

लेख के शेष भाग का ढांचा इस प्रकार है। भाग II में संबंधित साहित्य की समीक्षा की गई है। भाग III में मनोभाव वर्गीकरण की कार्य पद्धति और मनोभाव सूचकांक के निर्माण का वर्णन किया गया है। भाग IV में मीडिया के मनोभाव एवं मुद्रास्फीति के बीच के संबंध को परखने वाले अनुभवजन्य परिणाम प्रस्तुत किए गए हैं और भाग V में निष्कर्ष दिए गए हैं।

II. साहित्य की समीक्षा

समाचारपत्र सूचना आधारित सूचकांकों का निर्माण किया गया और उसका विभिन्न समष्टि-आर्थिक एवं वित्तीय विश्लेषण में उपयोग किया गया, जैसे आर्थिक नीति अनिश्चितता (बेकर और अन्य, 2016; भगत और अन्य, 2013), वित्तीय बाजार संचलन (बेकर और अन्य, 2019; मनीला और मोरीरा, 2017), समष्टि-आर्थिक चरों का अपेक्षित विकास (बेकर्स और अन्य, 2017; शपिरो और अन्य, 2017) तथा केंद्रीय बैंक से जुड़ी संभाव्य नीतिगत प्रतिक्रिया (लैमला और स्टर्म, 2013; हेन्ड्री, 2012; टोब्बैक और अन्य, 2017)। हम मुद्रास्फीति के विश्लेषण से जुड़े कतिपय अध्ययनों की संक्षेप में समीक्षा करेंगे।

मीडिया द्वारा प्रदान की गई सूचना को लेकर उपभोक्ताओं की प्रतिक्रिया को इस साहित्य में बखूबी परखा गया है। आम जनता को समष्टि-आर्थिक मॉडल की पूरी समझ नहीं हो सकती है, और साथ ही वे ताजे आंकड़ों को ट्रैक नहीं कर सकते हैं, एवं इसके बजाय वे समष्टि-आर्थिक गतिविधियों पर ताजी जानकारियों के लिए न्यूज़ मीडिया पर निर्भर हो सकते हैं ताकि वे अपने

---

* यह लेख भारतीय रिज़र्व बैंक (आरबीआई) के सांख्यिकी और सूचना प्रबंध विभाग (डीएसआईएम) की श्वेता कुमारी और गीता गिड्डी द्वारा तैयार किया गया है। इस लेख में व्यक्त विचार लेखकों के अपने विचार हैं और वे बैंक के विचारों को नहीं दर्शाते हैं। यदि कोई त्रुटियां होती हैं, तो वे लेखकों के हैं।

पूर्वानुमान और अपनी अपेक्षाएं बना सकें। लिहाजा, समाचार के प्रेषक के तौर पर मीडिया का परिवारों की मुद्रास्फीति संबंधी अपेक्षाओं पर सीधा प्रभाव पड़ सकता है (कैरोल, 2003)। न्यूज़ मीडिया द्वारा प्रदान की गई सूचनाओं को लेकर उपभोक्ताओं की प्रतिक्रिया पर मात्रा (न्यूज़ कवरेज) और गुणवत्ता (समाचार का लहजा) दोनों का प्रभाव पड़ सकता है (लैमला और लेइन, 2008)।

उपभोक्ताओं की मुद्रास्फीति संबंधी अपेक्षाओं को लेकर मीडिया न्यूज़ में विविधता पाई जाती है, जहां रिपोर्टिंग की तीव्रता और समाचार की विषय-वस्तु बड़ी अहम भूमिका अदा कर सकती है (लैमला और माग, 2012)। मीडिया को, विभिन्न समष्टि-आर्थिक विशेषताओं के अतिरिक्त, एक महत्वपूर्ण प्रभावकर्ता के रूप में पाया जाता है जो मुद्रास्फीति संबंधी अपेक्षाओं के संभाव्य निर्धारक के रूप में कार्य करती है (एहरमन और अन्य, 2017)। इससे थोड़ा भिन्न रास्ता अपनाते हैं तो पता चलता है कि कतिपय अध्ययनों में समाचारों के मनोभाव और कारोबार चक्र के संकेतकों के बीच की कड़ी पर ध्यान केंद्रित किया गया है। यह पाया गया है कि समाचारों के मनोभाव के उपयोग से मॉडल के पूर्वानुमान करने की क्षमता में सुधार होता है (बेकर्स और अन्य, 2017; शापिरो और अन्य, 2017)।

III. मनोभाव वर्गीकरण और मनोभाव सूचकांक की कार्य पद्धति

ऑनलाइन डेटा स्रोत, समाचारों का लाभ उठाने का एक अवसर प्रदान करता है, जो अधिक तादाद एवं अव्यवस्थित पाठ्य स्वरूप में होते हैं, और जो इन्सान द्वारा पठन एवं प्रोसेसिंग को चुनौतीपूर्ण बनाता है। इस रचना में, मनोभाव के विश्लेषण के लिए तीन व्यापक दृष्टिकोण अपनाए गए हैं, जिसके लिए शब्दकोष आधारित दृष्टिकोण, अर्थ विन्यास और मशीन लर्निंग तकनीकों जैसे कच्चे समाचार पाठों का उपयोग किया गया है।

मनोभाव वर्गीकरण के लिए शब्दकोष आधारित पद्धतियों (जैसे, लॉगरन-मेकडॉनल्ड शब्दकोष) को समझने और लागू करने में आसानी होती है, जिसका प्रयोग अर्थशास्त्र और वित्त (इग्लेसियस और अन्य, 2017; नीमन और अन्य, 2018) में किया गया है। जहां इस तरह की पद्धतियों से विषय-वस्तु में निहित आम मनोभाव को निकालने में मदद मिलती हैं, वहीं वे संदर्भ विशिष्ट मनोभाव के लिए उपयुक्त नहीं हो सकतीं क्योंकि वे सामान्य स्वरूप की होती हैं और किसी अमुक संदर्भ (जैसे, मुद्रास्फीति) के लिए सुपरिभाषित नहीं हैं।

अर्थ विन्यास (एसओ) दृष्टिकोण, विशिष्ट संदर्भ के मुद्दे का समाधान करने का प्रयास करता है, क्योंकि शोधकर्ता बहिर्जात रूप से पूर्व-निर्धारित संकेतशब्दों की सूची प्रदान कर सकता है जो किसी अमुक संदर्भ के लिए उपयुक्त माना जाता है। एसओ दृष्टिकोण, उपयोगकर्ता द्वारा प्रदान किए गए संकेतशब्दों का अनुसरण करते हुए, दिए गए पाठ में सकारात्मकता या नकारात्मकता के स्तर की माप करने का लक्ष्य रखता है (लुक्का और ट्रैबी, 2009; टर्नी, 2002; टोब्बैक और अन्य, 2017)। एसओ दृष्टिकोण सीधा-सादा है तथा आसानी से अपनाया जा सकता है; फिर भी उसकी कुछ सीमाएं हैं, जैसे कि संकेतशब्दों पर अत्यधिक निर्भरता और पक्षपात की संभावना (टोब्बैक और अन्य, 2017)।

मशीन लर्निंग (एमएल) पद्धतियों का उपयोग करके शोधकर्ता द्वारा बहिर्जात रूप से उपयुक्त संकेतशब्द प्रदान करने के मुद्दे का समाधान किया जाता है। ये पद्धतियां दिए गए पाठ्य दस्तावेजों में शब्दों/स्वरूप को स्वतः खोजती हैं जो एक मनोभाव वर्ग को दूसरे से अलग करता है, और हाल की अवधियों में उपयोग की जा रही हैं (टोब्बैक और अन्य, 2017; शापिरो और अन्य, 2017)। एसवीएम एक संचालित मशीन लर्निंग तकनीक है जो मनोभाव वर्गीकरण के लिए व्यापक रूप से उपयोग की गई पद्धति है।

III.1 मनोभाव वर्गीकरण - कार्य पद्धति

हम ऑनलाइन प्रिंट मीडिया से भारत में खुदरा मुद्रास्फीति पर केंद्रित समाचार जुटाते हैं। सूचना अपकर्षण और मनोभाव वर्गीकरण कार्य पद्धतियां अंग्रेजी भाषा के लिए काफी हद तक विकसित की गई हैं, एवं इसलिए, हम अपने दायरे को इस लेख में अंग्रेजी समाचारों तक ही सीमित रखते हैं। कोई, अन्य भारतीय भाषाओं में भी समाचारों से मनोभाव पता लगाने एवं निकालने की सोच रख सकता है, ताकि भाषा (यदि कोई) को लेकर मनोभाव में संभाव्य परिवर्तनों की जांच की जा सके। फिर भी, भिन्न-भिन्न राज्यों/क्षेत्रों में प्रिंट मीडिया में अंग्रेजी के व्यापक प्रयोग के चलते, हमारा मानना है कि अंग्रेजी में लिखे गए समाचारों से निकाले गए मनोभाव पर्याप्त मात्रा में प्रतिनिधित्व करेंगे।

जहां एसवीएम क्लासिफायर एक संचालित मॉडल[1] है, वहीं विभिन्न विशेषताओं के साथ लेबल किए गए दस्तावेज और

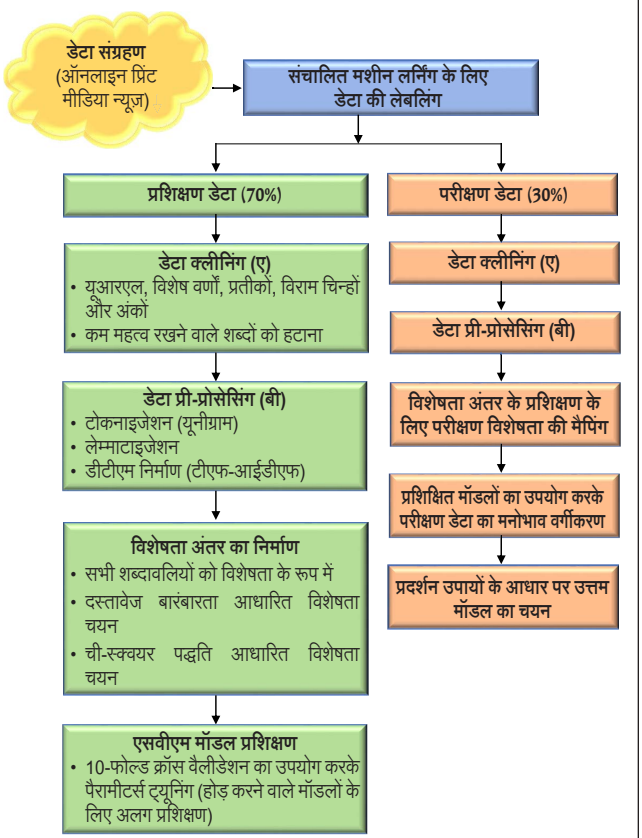
[1] इस लेख में एसवीएम की अत्यंत संक्षिप्त व्याख्या की गई है। विवरण के लिए चक्रबर्ती और जोसफ, 2017 का संदर्भ लिया जा सकता है।

दस्तावेज-शब्द मैट्रिक्स (डीटीएम) के रूप में इनपुट, मॉडल[2] को प्रशिक्षित करने की पूर्व-आवश्यकताएं हैं। वर्तमान मामले में, प्रत्येक समाचार को अद्वितीय दस्तावेज के रूप में माना गया और समाचारों में उल्लेख किए गए शब्दों को शब्दावलियों के रूप में प्रयोग किया गया।

समाचार पाठ को भली-भांति पढ़कर प्रत्येक समाचार (वृद्धि, कमी या तटस्थ) को मनोभाव सौंपा गया था। कुछ समाचारों को मनोभाव के तीन वर्गों (अर्थात, मुद्रास्फीति को लेकर कोई मनोभाव नहीं था) में से किसी में वर्गीकृत नहीं किया जा सका, एवं इसलिए 'शून्य' के रूप में लेबल किया गया था। इस प्रक्रिया के परिणामस्वरूप सभी दस्तावेजों को चार वर्गों/श्रेणियों, अर्थात, 'वृद्धि', 'कमी', 'तटस्थ' या 'शून्य' में से एक में लेबल किया गया।

हम डेटा संग्रहण से लेकर मनोभाव वर्गीकरण की समूची प्रक्रिया का वर्णन एक फ्लो चार्ट (चार्ट 1) के माध्यम से करते हैं, जिसका विवरण अनुबंध I में दिया गया है।

चार्ट 1: फ्लो चार्ट – मनोभाव वर्गीकरण



[2] हम एसवीएम मॉडल प्रशिक्षण और परीक्षण के लिए, आर में सीएआरईटी पैकेज का उपयोग करते हैं।

एसवीएम में उपयोग की गई अंतर्निहित अवधारणा विशिष्ट लक्षणों की पहचान करने के लिए है जिसका उपयोग एक मनोभाव वर्ग को दूसरे से अलग करने के लिए किया जा सकता है, एवं इसलिए, अफवाह को कम करके तथा वर्गीकरण मॉडल की सटीकता को बेहतर बनाकर विशेषताओं के सही मिश्रण वाले कुछ एक दस्तावेजों को काम में लाया जा सकता है।

अतः विशिष्ट लक्षणों को चुनने के लिए मनोभाव विश्लेषण में विशेषता चयन पद्धतियों का प्रायः उपयोग किया जाता है जो तुलनात्मक रूप से अधिक सूचनाप्रद होते हैं और उच्चतर वर्गीकरण, सटीकता हासिल करने में सहायक होगी। मूल विचार यह है कि विशेषताओं को निर्धारित उपायों के अनुसार श्रेणीबद्ध किया जाए और गैर-सूचनाप्रद विशेषताओं को हटाया जाए। हमने दो उपायों की खोज की है, जैसे, विशेषता चयन के लिए दस्तावेज बारंबारता आधारित और ची-स्क्वयर दृष्टिकोण। विवरण अनुबंध II में उपलब्ध कराए गए हैं। लिहाजा, सभी विशेषताओं वाले आधारभूत मॉडल के अतिरिक्त, दो विशेषता चयन पद्धतियों के अनुरूप दो और प्रकार के मॉडल को प्रशिक्षित किया गया था।

जैसा कि मशीन लर्निंग एल्गोरिदम का सामान्य दस्तूर है, मॉडल के प्रदर्शन को परखने के लिए डेटा को प्रशिक्षण और परीक्षण डेटा के रूप में अलग किया गया तथा परीक्षण डेटा में पर्फार्मन्स मेट्रिक नामतः 'सटीकता अनुपात' के आधार पर होड़ में लगे विभिन्न मॉडलों में से इष्टतम मॉडल का चयन किया गया। सटीकता अनुपात, वर्गीकरण मॉडलों के लिए एक मानक मूल्यांकन मेट्रिक/उपाय है, जो कुल प्रतिक्रियाओं की तुलना में सही तरीके से वर्गीकृत प्रतिक्रियाओं के अनुपात को दर्शाता है। उच्च सटीकता अनुपात (परीक्षण डेटा में) वाला मॉडल होड़ में लगे मॉडलों में बेहतर माना जाता है।

कॉर्पस को प्रशिक्षण और परीक्षण डेटा सेट में 70:30 के अनुपात में अलग-अलग किया गया था, और सारणी I, प्रशिक्षण और परीक्षण डेटा से संबंधित समाचारों के अनुपात को दर्शाती है।

सारणी 1: समाचारों का विभाजन

लेबल	प्रशिक्षण	परीक्षण	कुल
कमी	962	411	1373
वृद्धि	755	324	1079
तटस्थ	15	7	22
शून्य	3744	1605	5349
कुल	**5476**	**2347**	**7823**

जैसा सारणी 1 में देखा जा सकता है, सिर्फ कुछ एक ही 'तटस्थ' वर्ग में आते हैं, जिससे डेटा को लेकर असंतुलन की समस्या पैदा होती है। गुणात्मक सर्वेक्षण परिणामों (सुस्पष्ट डेटा) में यह आम बात है और साथ ही उन मामलों में जहां लक्ष्य चर लगातार बदल रहा है (दोनों दिशाओं में) एवं चर की समान स्थिति में बने रहने की संभावना बहुत ही कम है।

असंतुलित डेटा के साथ पेश आने का एक दृष्टिकोण है कि निम्न ऑबज़र्वेशन वर्ग को दूसरे (बगल) वर्ग के साथ मिलाना, कोई 'तटस्थ' वर्ग को या तो 'वृद्धि' अथवा 'कमी' के साथ मिलाने के बारे में सोच सकता है। तथापि, इस लेख में, इस दृष्टिकोण पर विचार नहीं किया गया, क्योंकि मनोभाव सूचकांक (बाद के चरण में) विकसित करने के लिए तीन वर्गों ('वृद्धि', 'कमी' और 'तटस्थ') की जरूरत होती है।

जैसा पहले वर्णन किया गया है, भिन्न-भिन्न विशेषता अंतर के साथ अलग-अलग मॉडलों का प्रशिक्षण और परीक्षण किया गया (सारणी 2)। दूसरे मॉडल में, भले ही विशेषताओं की संख्या पहले मॉडल की अपेक्षा निश्चय ही बहुत कम है, लेकिन सटीकता काफी बेहतर हुई है, जो इस तथ्य पर प्रकाश डालती है कि उपयुक्त विशेषताओं के साथ व्यवहार करने से बेहतर परिणाम आते हैं। इसलिए, मनोभाव वर्गीकरण के लिए दूसरे मॉडल को इष्टतम मॉडल के रूप में चुना गया क्योंकि परीक्षण डेटा में उसकी सटीकता अन्य मॉडलों की तुलना में अधिक थी। मॉडल 2 का उपयोग करके अप्रैल 2015 से मार्च 2019 तक सभी समाचारों के लिए मनोभाव सौंपे गए।

III.2 मनोभाव सूचकांक - कार्य पद्धति

दस्तावेजों (समाचारों) को चार मनोभाव वर्गों, अर्थात, 'वृद्धि', 'कमी', 'तटस्थ' और 'शून्य' में वर्गीकृत करने के उपरांत अगला कदम उनको एकत्रित करना होता है और फिर प्रत्येक समयावधि के लिए समग्र मनोभाव निकालना होता है। मुद्रास्फीति एक समयावधि से दूसरी में बदलती है (कम या ज्यादा) और लगातार दो समयावधियों में समान रहने की उसकी संभावना निश्चय ही बहुत कम होती है। अतः, 'तटस्थ' मनोभाव वाले समाचारों की संभावना निश्चय ही बहुत कम होने की उम्मीद है। फिर भी, मनोभाव के सभी पहलुओं को कवर करने के लिए हम इस तरह के सभी समाचारों को कॉर्पस में गिनते हैं। 'शून्य' के रूप में वर्गीकृत समाचारों को मनोभाव सूचकांक की गणना से हटा दिया जाता है क्योंकि वे कोई मनोभाव नहीं सूचित करते हैं और उनको शामिल करने से सूचकांक अशुद्ध हो सकता है।

सारणी 2: एसवीएम मॉडल के विभिन्न प्रकारों के प्रदर्शन के परिणाम

मॉडल	मॉडल	विशेषताओं की संख्या	प्रशिक्षण – सटीकता (प्रतिशत में)	परीक्षण – सटीकता (प्रतिशत में)
1	सभी शब्दावलियों को विशेषताओं के रूप में उपयोग करना	2574	99	64
2	दस्तावेज बारंबारता पद्धति का उपयोग करके विशेषता का चयन	186	92	90
3	ची-स्क्वयर पद्धति का उपयोग करके विशेषता का चयन	178	92	61

समाचारों के मनोभाव को मापना कुछ हद तक गुणात्मक कारोबार प्रवृत्ति सर्वेक्षणों के अनुरूप है, जो सर्वेक्षण के चिन्हित प्रतिवादियों के मनोभाव/अपेक्षाओं को मापने का लक्ष्य रखता है। सर्वेक्षण की प्रतिक्रियाओं को आम तौर पर शुद्ध प्रतिक्रिया का उपयोग करके एकत्रित किया जाता है, जो 'वृद्धि' एवं 'कमी' प्रतिक्रियाओं (अनुपात) के बीच का अंतर होता है। हम निम्नानुसार मनोभाव सूचकांक (एसआई) विकसित करते हैं :

$$SI_t = \left(\frac{I_t - D_t}{N_t}\right) X\ 100 \qquad \ldots (1)$$

जहां I_t = समयावधि t में 'वृद्धि' मनोभाव वाले समाचारों की संख्या

D_t = समयावधि t में 'कमी' मनोभाव वाले समाचारों की संख्या

N_t = कुल समाचारों की संख्या (वृद्धि, कमी और तटस्थ)

एसआई (-)100 से (+)100 के बीच रहता है, जहां सूचकांक का ऋणात्मक मान मुद्रास्फीति में कमी और धनात्मक मान मुद्रास्फीति में वृद्धि दर्शाता है।

समाचारों की दैनिक उपलब्धता से एसआई की दैनिक आधार पर गणना में सुविधा होती है, जो इसे अधिक दोहराने वाला संकेतक बनाता है। फिर भी, आधिकारिक मासिक मुद्रास्फीति संख्याओं की तुलना में आकलन के प्रयोजन से हम प्रत्येक माह के लिए एसआई की गणना करते हैं, जिसके लिए माह के सभी दिनों को हिसाब में लिया जाता है।

साथ ही, हम दैनिक समाचारों में निहित महत्वपूर्ण जानकारियों को खोना नहीं चाहते हैं। इसलिए, हम माह के कतिपय दिनों को जोड़ने और एसआई की फिर से गणना करने का प्रयास करते हैं। मुद्रास्फीति के बारे में मीडिया में अक्सर रिपोर्ट आती है

क्योंकि वह एक ऐसा समष्टि-अर्थिक चर है जिसे बड़ी उत्सुकता के साथ देखा जाता है। तथापि, उपभोक्ता मूल्य सूचकांक (सीपीआई) मुद्रास्फीति पर आधिकारिक डेटा रिलीज करने के दौरान उसका कवरेज अधिक बढ़ जाता है।

इस पृष्ठभूमि में, हम प्रत्येक माह के दिनों को तीन अलग सेट्स में मिलाते हैं :

सेट 1 : दिन 1 से टी-1

सेट 2 : दिन टी से टी + 2

सेट 3 : माह के शेष दिन।

जहां टी = मासिक सीपीआई डेटा पर प्रेस-रिलीज जारी करने की तारीख, जिसमें भिन्न-भिन्न माहों में कभी-कभी अंतर हो सकता है।

अतः, हम प्रत्येक माह के लिए चार एसआई मूल्यों की गणना करते हैं, प्रत्येक तारीख सेट (सेट 1, सेट 2 और सेट 3) के लिए एक और माह के सभी दिनों समेत समग्र मासिक सूचकांक। लक्ष्य यह है कि समाचार प्राप्त होने के समय के कारण मुद्रास्फीति संबंधी मनोभाव में संभाव्य अंतर, यदि कोई हो, का पता लगाया जा सके। समाचारों को सेट्स के रूप में वर्गीकृत करने से निम्नलिखित पहलुओं में फायदा होता है :

(i) कम गणना करना पड़ता है, समय-विशिष्ट मनोभाव;

(ii) मनोभाव, माह समाप्त होने से काफी पहले ही उपलब्ध हो जाना; और

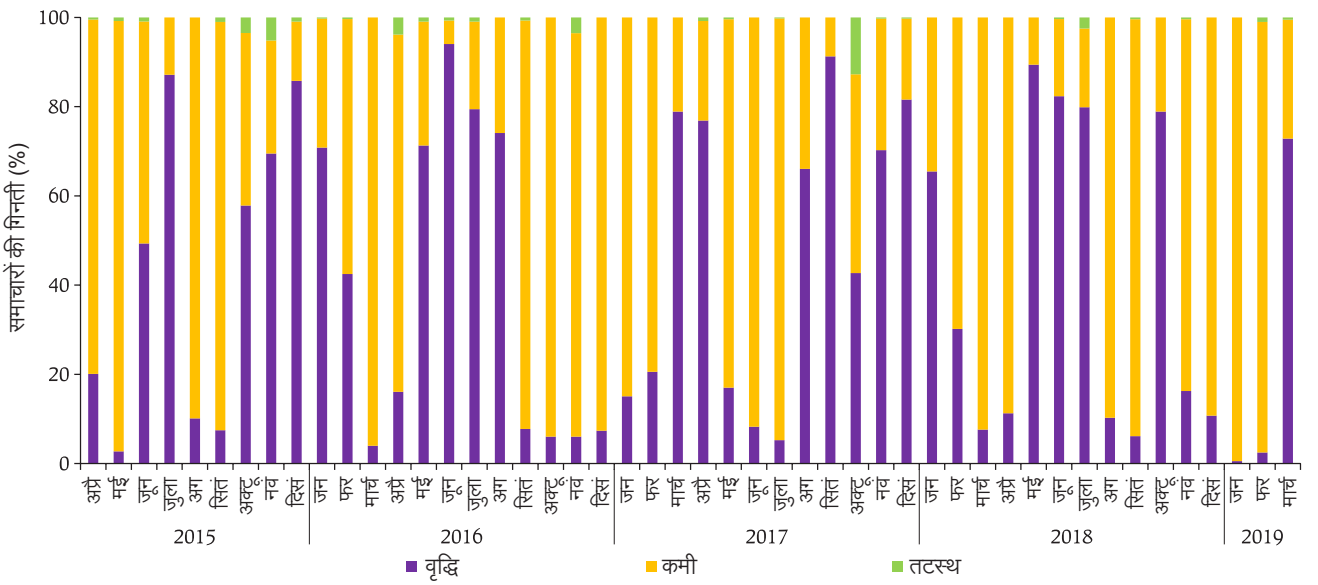
(iii) यदि किसी अमुक सेट का मनोभाव, मुद्रास्फीति के साथ बेहतर रूप से जुड़ा है तो हमें अन्य दिनों पर सोच-विचार करने की जरूरत नहीं है, भले ही वह फायदेमंद हो।

III.3 मनोभाव सूचकांक पर शोधपरक तथ्य

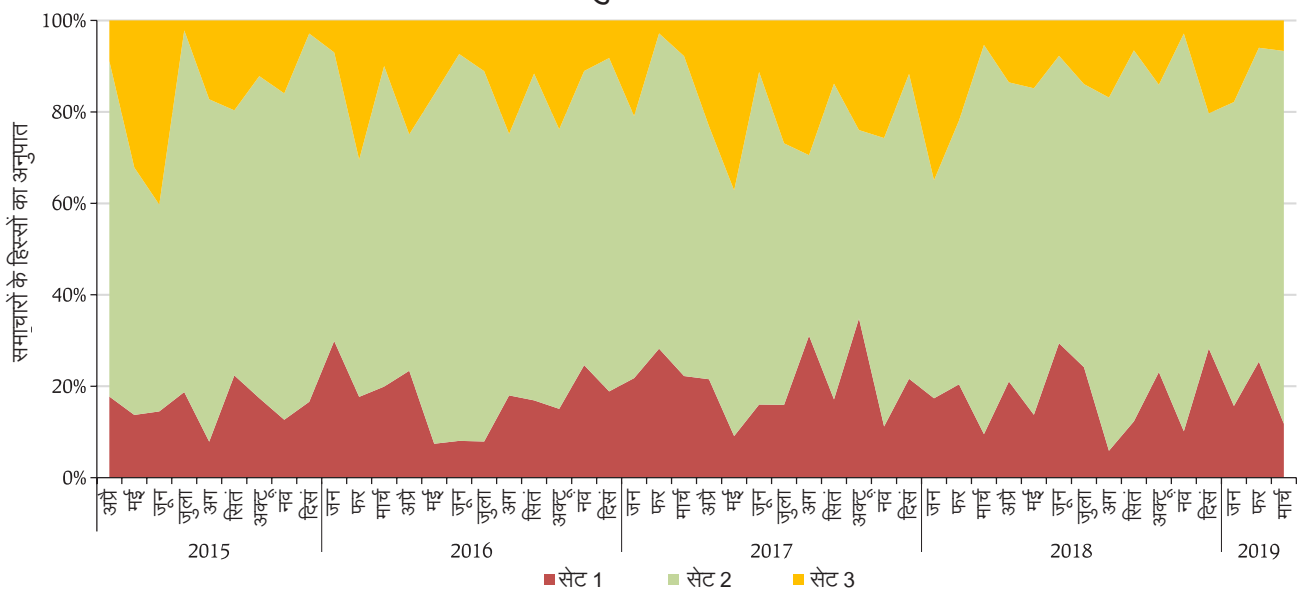
इस भाग में, मनोभाव सूचकांक के विशेष लक्षणों का कुल एवं भिन्न-भिन्न स्तर पर वर्णन किया गया है। चार्ट 2 में प्रत्येक माह में मनोभाव के आधार पर समाचारों के विभाजन को दर्शाया गया है। यह देखा जा सकता है कि 'तटस्थ' के रूप में वर्गीकृत समाचारों का अनुपात करीब-करीब अधिकांश महीनों के लिए नगण्य है (यह अपेक्षा के अनुरूप है, इस तथ्य को देखते हुए कि लक्ष्य चर, अर्थात मुद्रास्फीति लगातार बदल रही है)। इसके अलावा, प्रत्येक महीने में, समाचार का अधिकांश हिस्सा मुख्य रूप से 'वृद्धि' या 'कमी' मनोभाव वर्ग पर केंद्रित होता है, और ऐसे उदाहरण दुर्लभ होते हैं जहां दोनों प्रकार के मनोभाव मीडिया में समान रूप से प्रचलित हैं। मीडिया के मनोभाव के ऐसे संकेद्रण से मनोभाव निर्माण में स्पष्टता का पता चलता है, और नतीजतन, मनोभाव सूचकांक का मूल्य, अधिक ऋणात्मक अथवा अधिक धनात्मक होता है।

तीन तारीख सेटों के संबंध में समाचारों के विभाजन को चार्ट 3 में प्रस्तुत किया गया है।

चार्ट 2: मनोभाव के अनुसार समाचारों का विभाजन



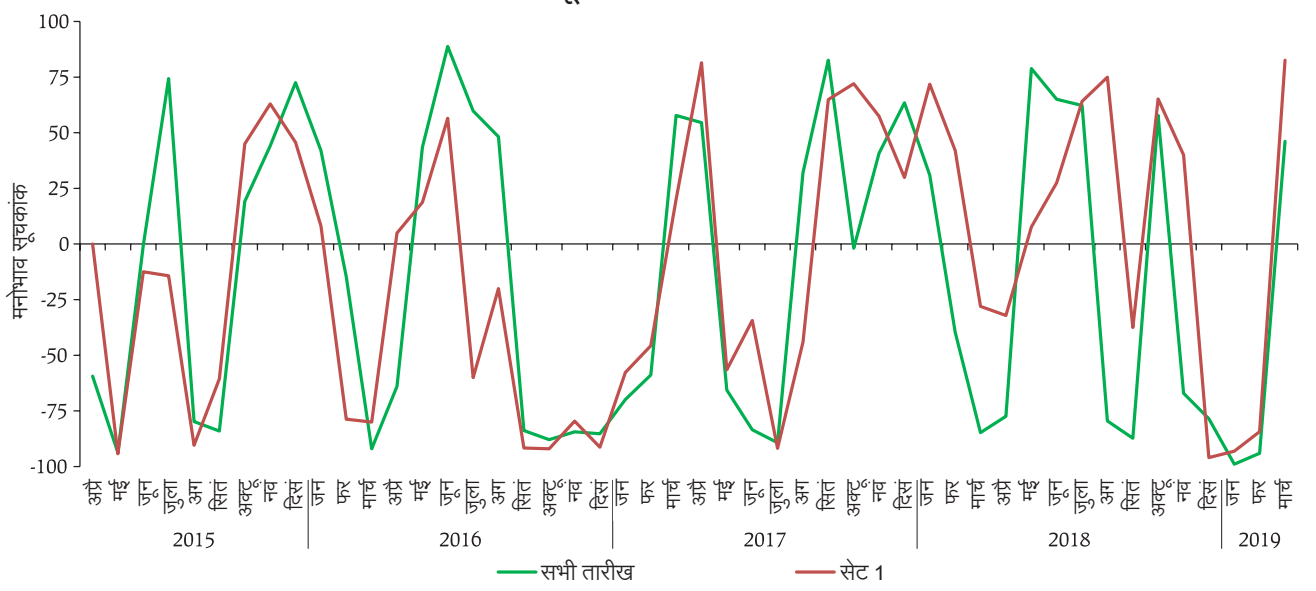
चार्ट 3: सेट के अनुसार समाचारों का विभाजन



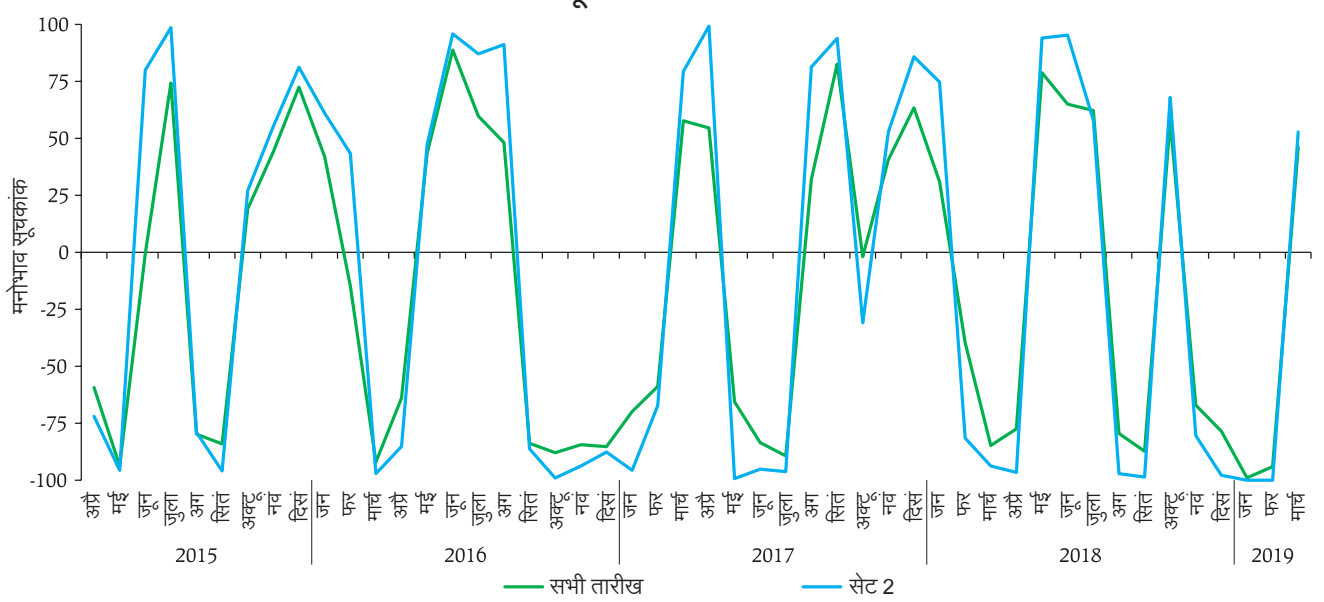
यह देखा जा सकता है कि यद्यपि सेट 2 में महीने के सिर्फ तीन दिन होते है, लेकिन अधिकांश समाचार इसी अवधि के दौरान प्राप्त किए जाते हैं। यह भली-भांति समझ में आता है कि मौजूदा मुद्रास्फीति के रुझानों एवं संकेतक के संभावित भावी राह को लेकर बहुत सारे विचार-विमर्श मुद्रास्फीति के आंकड़ों की आधिकारिक रिलीज की तारीख के आसपास होते हैं।

यह देखना दिलचस्प होगा कि माह के समग्र मनोभाव में सेट 2 के दौरान कवर किए गए अधिकांश समाचारों की क्या कोई भूमिका है। हम प्रत्येक सेट के सम्मुख समग्र मनोभाव सूचकांक को अंकित करते हैं (चार्ट 4 से चार्ट 6)। जहां, सेट 2 पर आधारित सूचकांक सभी तारीखों के सूचकांक को बेहतर दर्शाता है, वहीं सेट 1 और सेट 3 के सूचकांक के मामले में ऐसा

चार्ट 4: मनोभाव सूचकांक – सभी तारीख और सेट 1

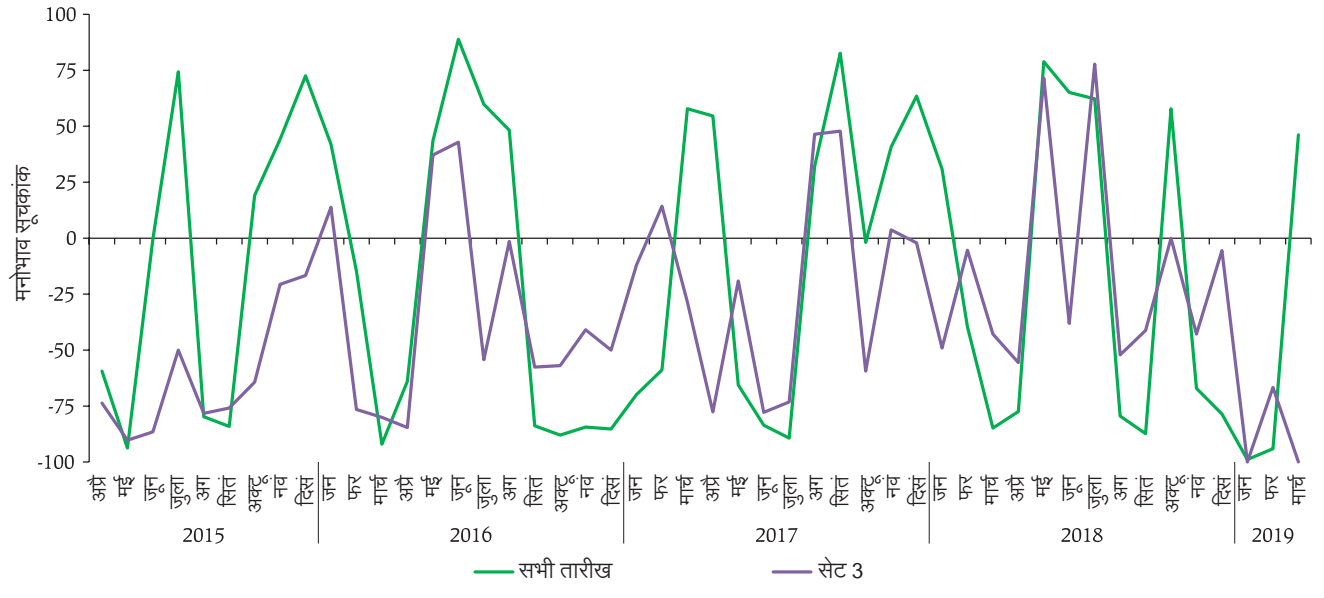


चार्ट 5: मनोभाव सूचकांक – सभी तारीख और सेट 2



नहीं पाया गया। हो सकता है कि सेट 2 में अधिकांश समाचारों को कवर किए जाने के साथ-साथ मनोभाव के बेहतर कवरेज की वजह से सेट 2 के सूचकांक और सभी तारीखों के सूचकांक के बीच सह-संचलन का स्तर बढ़ गया है।

चार्ट 6: मनोभाव सूचकांक – सभी तारीख और सेट 3



IV. प्रयोगसिद्ध विश्लेषण

निर्मित मनोभाव सूचकांक, आधिकारिक सीपीआई डेटा जारी होने से लगभग एक पखवाड़े पहले उपलब्ध होता है। इसके अतिरिक्त, सेट 2 से संबंधित मनोभाव सूचकांक, मुद्रास्फीति पर डेटा जारी होने से लगभग एक महीने पहले उपलब्ध होता है। मनोभाव प्राप्त करने के लिए अंग्रेजी समाचारों पर विचार किया जाता है, जिसे ज्यादातर देश के शहरी हिस्सों में पढ़ा जा सकता है, इसलिए मुमकिन है कि मनोभाव सूचकांक शहरी मुद्रास्फीति से बेहतर संबंधित है। दूसरी तरफ, चूंकि समाचार पत्र विभिन्न भाषाओं में आमतौर पर किसी भी समय समान मुद्दों के बारे में रिपोर्ट करते हैं, इसलिए सूचकांक को ग्रामीण मुद्रास्फीति के साथ भी जोड़ा जा सकता है। इसलिए, हम इस लेख में संयुक्त, शहरी और ग्रामीण मुद्रास्फीति पर विचार करते हैं।

हम मुद्रास्फीति को सीपीआई के लॉगरिद्मिक मूल्यों में वार्षिक परिवर्तन के रूप में परिभाषित करते हैं। मनोभाव सूचकांक, मुद्रास्फीति में परिवर्तन की दिशा को इंगित करता है, इसलिए हम मुद्रास्फीति में मासिक परिवर्तन को निम्नानुसार परिभाषित करते हैं:

$$\pi_{i,t} = (\log_e CPI_{i,t} - \log_e CPI_{i,t-12}) \times 100 \quad \ldots.(2)$$

$$\Delta\pi_{i,t} = \pi_{i,t} - \pi_{i,t-1} \quad \ldots.(3)$$

जहां $CPI_{i,t}$ = समय t में वर्ग i का CPI

(i = C संयुक्त के लिए, i = U शहरी के लिए, i = R ग्रामीण के लिए)

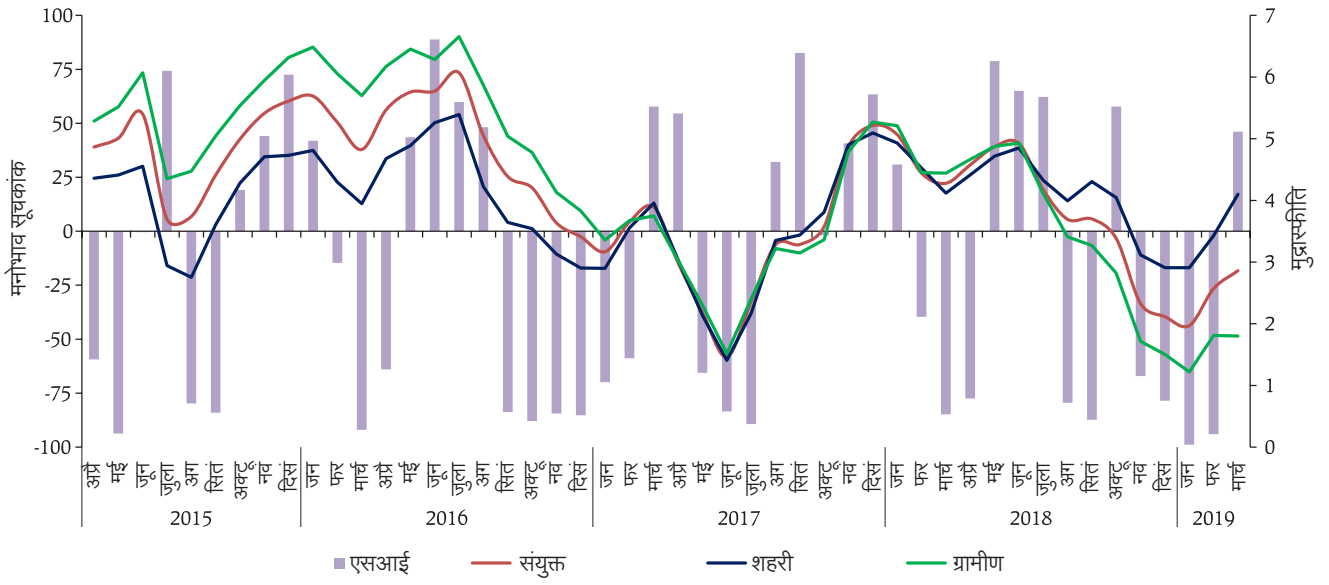
वर्तमान विश्लेषण में अप्रैल 2015 से मार्च 2019 तक के आंकड़ों पर विचार किया गया है। हम कारण-कार्य संबंध के परीक्षण से पहले रेखा-चित्रीय, सहसंबंध और दिशात्मक विश्लेषण से प्रारंभ करते हैं।

IV.1 रेखा-चित्रीय और सहसंबंध विश्लेषण

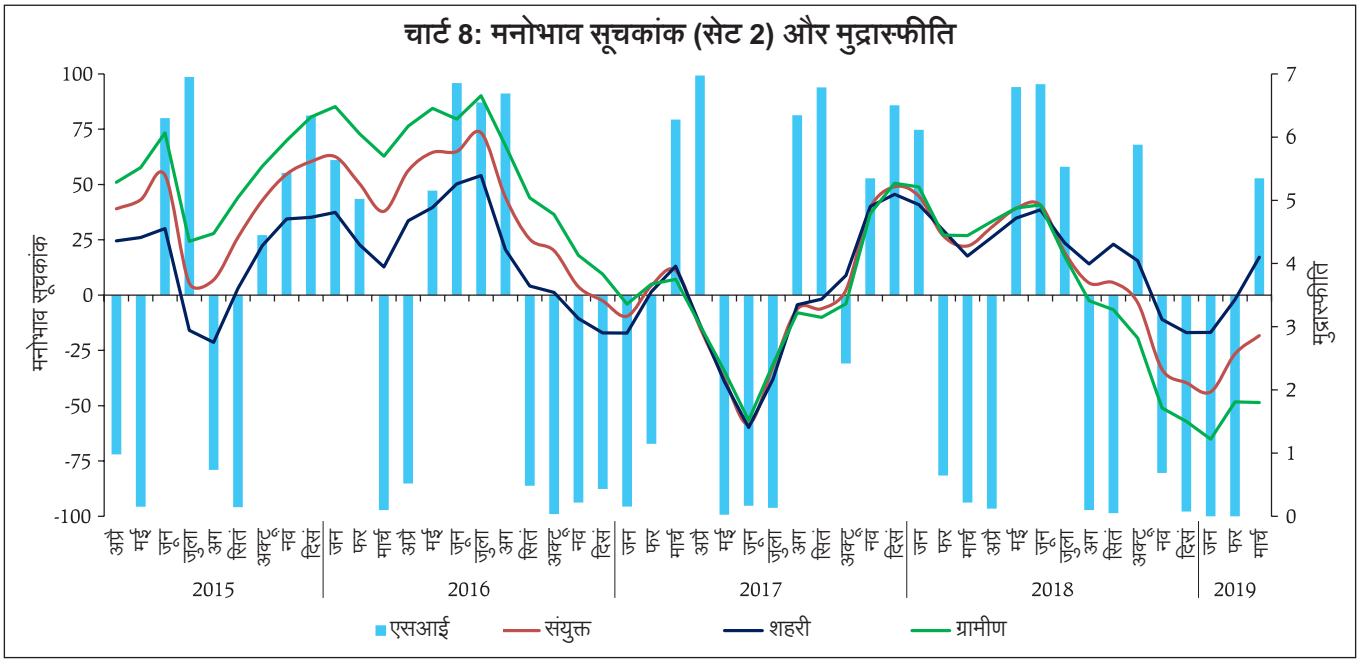
मनोभाव सूचकांक और मुद्रास्फीति के बीच उच्च स्तर का सह-संचलन पाया गया है, जो दर्शाता है कि मनोभाव सूचकांक मुद्रास्फीति में दिशात्मक परिवर्तन को भली-भांति ट्रैक करने के योग्य है (चार्ट 7 और चार्ट 8)।

सहसंबंध विश्लेषण से पता चलता है कि मनोभाव और मुद्रास्फीति के बीच दृढ़ और महत्वपूर्ण सहसंबंध है। यह सहसंबंध शहरी मुद्रास्फीति की तुलना में ग्रामीण मुद्रास्फीति के लिए कमजोर होता है, लेकिन आंकड़ों की दृष्टि से महत्वपूर्ण है। सेट 2 के एसआई का मुद्रास्फीति के साथ मजूबत सहसंबंध है, जबकि सेट 1 और सेट 3 के मनोभाव सूचकांक और मुद्रास्फीति के बीच उल्लेखनीय रूप से सहसंबंध प्रतीत नहीं होता है (सारणी 3)।

चार्ट 7: मनोभाव सूचकांक (समग्र) और मुद्रास्फीति



चार्ट 8: मनोभाव सूचकांक (सेट 2) और मुद्रास्फीति



IV.2 दिशात्मक विश्लेषण

मुद्रास्फीति में टर्निंग पाइंट्स का पता लगाने में मनोभाव सूचकांक के ट्रैकिंग पर्फार्मन्स का आकलन करने के लिए दिशात्मक विश्लेषण को अपनाया जाता है। वैसे तो कई मानक दिशात्मक मेट्रिक्स होते हैं, लेकिन हम सटीकता का उपयोग करते हैं, जो इस रचना में बिलकुल सरल और व्यापक रूप से उपयोग किया गया मेट्रिक है। इसे सामान्य तौर पर दृष्टांत के अनुपात के रूप में परिभाषित किया जाता है, जब अनुमानित वर्ग संदर्भ वर्ग से मेल खाता है। हमारे मामले में, इसे समयावधि (महीनों) के अनुपात के रूप में परिभाषित किया जाता है, जब मनोभाव सूचकांक द्वारा इंगित दिशा, मुद्रास्फीति में परिवर्तन की दिशा से मेल खाता है।

धनात्मक एसआई मुद्रास्फीति में वृद्धि दर्शाता है, वहीं ऋणात्मक एसआई मुद्रास्फीति में कमी दर्शाता है। प्रत्येक माह के लिए एसआई और $\Delta\pi$ (मुद्रास्फीति में परिवर्तन) के चिन्ह को नोट किया जाता है, और 2 x 2 आकस्मिकता सारणी तैयार करने के लिए प्रत्येक दिशात्मक परिवर्तन जोड़ी हेतु महीनों की संख्या को गिना जाता है, जिसका उपयोग करके सटीकता मापी जाती है (सारणी 4)।

उच्च सटीकता अंक से तात्पर्य है कि एसआई, मुद्रास्फीति में दिशात्मक परिवर्तन को भली-भांति पता लगाने के योग्य है।

हम दिशात्मक सटीकता को परखने के लिए फिशर्स एग्ज़ैक्ट (एफई) टेस्ट का भी उपयोग करते हैं। आकस्मिकता सारणी (सारणी 4) का उपयोग करके, शून्य परिकल्पना की जांच की

सारणी 3: मुद्रास्फीति और मनोभाव के बीच सहसंबंध

	π_C	π_U	π_R
एसआई$_0$	0.43*** (0.00)	0.52*** (0.00)	0.35** (0.02)
एसआई$_1$	0.23 (0.11)	0.45*** (0.00)	0.11 (0.48)
एसआई$_2$	0.45*** (0.00)	0.51*** (0.00)	0.38** (0.01)
एसआई$_3$	0.17 (0.24)	0.26 (0.07)	0.12 (0.42)

टिप्पणी : p- कोष्ठकों में दिया गया मान।

***, **, * 1, 5 और 10 प्रतिशत स्तर पर उल्लेखनीय के द्योतक हैं।

एसआई$_0$, एसआई$_1$, एसआई$_2$ और एसआई$_3$ क्रमशः समग्र (सभी तारीख), सेट 1, सेट 2 और सेट 3 से संबंधित मनोभाव सूचकांक दर्शाते हैं। π_C, π_U, और π_R क्रमशः संयुक्त, शहरी और ग्रामीण मुद्रास्फीति का संकेत देते हैं।

सारणी 4: आकस्मिकता सारणी

महीनों की संख्या		**$\Delta\pi$ का चिन्ह**	
		वृद्धि	**कमी**
मनोभाव सूचकांक का चिन्ह	+ (धनात्मक)	A	B
	- (ऋणात्मक)	C	D

$$\text{सटीकता } \frac{(A+D)}{N} * 100$$

जहां, $N = A+B+C+D$

जाती है कि क्या मनोभाव सूचकांक द्वारा दी गई दिशा और मुद्रास्फीति में परिवर्तन की दिशा स्वतंत्र हैं। शून्य परिकल्पना के खारिज होने का तात्पर्य है कि एसआई, मुद्रास्फीति में परिवर्तन की दिशा का पता लगाने में उपयोगी है।

एफई टेस्ट में पाए गए आवश्यक स्तर को इस प्रकार परिभाषित किया गया है।

$$P = \frac{((A+B)!\ (C+D)!\ (A+C)!\ (B+D)!)}{(A!\ B!\ C!\ D!\ N!)} \quad \ldots (4)$$

जैसा पहले देखा गया था, सेट 2 एसआई का समग्र एसआई के अलावा मुद्रास्फीति के साथ नजदीकी संबंध है और इसलिए हम दिशात्मक विश्लेषण के लिए समग्र एसआई और सेट 2 एसआई पर विचार करते हैं।

दिशात्मक उपायों के नतीजों, जैसे, सटीकता और एफई टेस्ट को सारणी 5 में दर्शाया गया है।

65 प्रतिशत की सटीकता के साथ, यह प्रतीत होता है कि मनोभाव सूचकांक संयुक्त और शहरी मुद्रास्फीति दोनों के मामले में दिशात्मक परिवर्तन का पर्याप्त रूप से पता लगाता है। ग्रामीण मुद्रास्फीति के मामले में सटीकता तुलनात्मक रूप से कम रहती है। एफई टेस्ट, ग्रामीण मुद्रास्फीति के सिवाय मनोभाव और मुद्रास्फीति के बीच आवश्यक संबंध की पुनःपुष्टि करता है।

IV.3 कारण-कार्य संबंध का विश्लेषण

कारण-कार्य संबंध की मौजूदगी, अनुमान लगाने की क्षमता को परखने के लिए महत्वपूर्ण है। इस रचना में प्रायः ग्रेंजर कारण-कार्य संबंध परीक्षण का उपयोग किया गया है ताकि दो चरों के बीच के आकस्मिक संबंध की जांच की जा सके। अंतर्निहित परिकल्पना यह है कि एक चर का पिछड़ा हुआ मान दूसरे चर में बदलाव और इसके विपरीत की व्याख्या करता है।

सारणी 5: प्रदर्शन सटीकता उपाय (दिशा)

	$\Delta\pi_C$	$\Delta\pi_U$	$\Delta\pi_R$
एसआई$_0$			
सटीकता	65%	67%	60%
एफई परीक्षण p-मान	0.04	0.02	0.24
एसआई$_2$			
सटीकता	65%	67%	60%
एफई परीक्षण p-मान	0.05	0.02	0.25

जैसा पहले देखा गया था, सेट 2 एसआई का समग्र एसआई के अलावा मुद्रास्फीति के साथ नजदीकी संबंध है और इसलिए हम कारण-कार्य संबंध परीक्षण के लिए समग्र एसआई और सेट 2 एसआई पर विचार करते हैं।

हम ग्रेंजर कारण-कार्य संबंध परीक्षण के लिए निम्नलिखित समीकरण जोड़ियों का आकलन करते हैं :

$$\Delta\pi_{i,t} = a + \sum_{k=1}^{n} \alpha_k \Delta\pi_{i,t-k} + \sum_{k=1}^{n} \beta_k SI_{j,t-k} + \varepsilon_t \ldots . (5)$$

$$SI_{j,t} = b + \sum_{k=1}^{n} \gamma_k \Delta\pi_{i,t-k} + \sum_{k=1}^{n} \delta_k SI_{j,t-k} + \eta_t \ldots . (6)$$

जहां एसआई और $\Delta\pi$ मनोभाव सूचकांक और मुद्रास्फीति में परिवर्तन हैं, जैसा उपर्युक्त समीकरण (1) और (3) में परिभाषित किया गया है। अधोलेख i मुद्रास्फीति के प्रकार (संयुक्त, शहरी अथवा ग्रामीण) को इंगित करता है, जबकि अधोलेख j एसआई के प्रकार (0 और 2 क्रमशः समग्र और सेट 2 के लिए) को दर्शाता है।

β_k की शून्य परिकल्पना का परीक्षण किया गया जो संयुक्त रूप से शून्य हैं, और उसका खारिज होना पुष्टि करता है कि एसआई का ग्रेंजर-कारण $\Delta\pi$ होता है। इसी तरह, γ_k की शून्य परिकल्पना का खारिज होना, जो संयुक्त रूप से शून्य हैं, पुष्टि करता है कि $\Delta\pi$ का ग्रेंजर-कारण एसआई होता है। श्वार्ज सूचना मानदंड (एसआईसी) का उपयोग करके अंतराल चयन (n का मान) किया जाता है।

हम ग्रेंजर कारण-कार्य संबंध परीक्षण करने से पहले एसआई और $\Delta\pi$ दोनों चरों में यूनिट रूट की मौजूदगी का परीक्षण करते हैं, ताकि यह सुनिश्चित किया जा सके कि चर स्थिर हैं। ऑगमेंटेड डिकी फुलर टेस्ट (एडीएफ) और फिलिप्स-पेरोन टेस्ट (पीपी) का उपयोग किया जाता है, जहां शून्य परिकल्पना का खारिज होना चर की स्थिरता की पुष्टि करेगा। यूनिट रूट टेस्ट से पता चलता है कि मनोभाव सूचकांक और मुद्रास्फीति में परिवर्तन स्थिर हैं (सारणी 6)।

सारणी 6: यूनिट रूट टेस्ट्स

चर	ऑगमेंटेड डिकी फुलर टेस्ट	फिलिप्स-पेरोन टेस्ट	इंटीग्रेशन
एसआई$_0$	-3.85 (0.02)	-4.61 (0.00)	I(0)
एसआई$_2$	-6.36 (0.00)	-6.50 (0.00)	I(0)
$\Delta\pi_C$	-5.15 (0.00)	-4.88 (0.00)	I(0)
$\Delta\pi_U$	-2.45 (0.35)	-4.42 (0.00)	I(0)
$\Delta\pi_R$	-5.33 (0.00)	-5.24 (0.00)	I(0)

टिप्पणी : p- कोष्ठकों में दिया गया मान ।

सारणी 7: ग्रेंजर कारण-कार्य संबंध परीक्षण के परिणाम

	$\Delta\pi_C$	$\Delta\pi_U$	$\Delta\pi_R$
एसआई$_0$			
एसआई, $\Delta\pi$ का ग्रेंजर कारण नहीं है	7.9398*** (0.0072)	11.3400*** (0.0015)	4.7973** (0.0338)
$\Delta\pi$, एसआई का ग्रेंजर कारण नहीं है	95.5920*** (0.0000)	99.6980*** (0.0000)	76.2310*** (0.0000)
एसआई$_2$			
एसआई, $\Delta\pi$ का ग्रेंजर कारण नहीं है	13.4300*** (0.0006)	16.6830*** (0.0001)	9.1198*** (0.0041)
$\Delta\pi$, एसआई का ग्रेंजर कारण नहीं है	84.5010*** (0.0000)	84.3350*** (0.0000)	70.3110*** (0.0000)

टिप्पणी : p- कोष्ठकों में दिया गया मान ।

***, **, * 1, 5 और 10 प्रतिशत स्तर पर उल्लेखनीय के द्योतक हैं।

ग्रेंजर कारण-कार्य संबंध परीक्षण यह दर्शाता है कि मनोभाव और मुद्रास्फीति में परिवर्तन (सारणी 7) के बीच द्वि-दिशात्मक कारण-कार्य मौजूद है (सारणी 7)।

ग्रेंजर कारण-कार्य संबंध परीक्षण के परिणामों से यह स्पष्ट है कि मुद्रास्फीति, अपने स्वयं के अंतराल के अलावा, मनोभाव से प्रभावित होती है और इसके विपरीत भी। इस प्रकार मनोभाव सूचकांक में मुद्रास्फीति का अनुमान लगाने के लिए आवश्यक व्याख्यात्मक सामर्थ्य है।

V. निष्कर्ष

इस लेख में मीडिया में रिपोर्ट की गई अत्यधिक दोहराई जाने वाली अव्यवस्थित सूचनाओं का उपयोग किया गया है और बिग डेटा तकनीकों का प्रयोग किया गया है ताकि निम्नलिखित उद्देश्य के साथ मनोभाव सूचकांक का निर्माण किया जा सके (ए) वैकल्पिक संकेतकों का निर्माण करने, जो लगभग तत्काल आधार पर अर्थव्यवस्था की स्थिति का आकलन करने के लिए उपयोगी हो सकता है, और (बी) मीडिया से प्राप्त जानकारी का उपयोग कर के मुद्रास्फीति के तात्कालिक अनुमान को बेहतर करना।

मशीन लर्निंग और नैचुरल लैंग्वेज प्रोसेसिंग तकनीकों, विशेष रूप से एसवीएम क्लासिफायर का उपयोग करते हुए, अव्यवस्थित पाठ (समाचार) से मनोभाव निकाले जाते हैं ताकि मनोभाव सूचकांक का निर्माण किया जा सके। अनुभवजन्य परिणाम बताते हैं कि मीडिया के मनोभाव सूचकांक मुद्रास्फीति को बहुत अच्छी तरह से ट्रैक करता है। इसकी दिशात्मक सटीकता, उच्च और आंकड़ों की दृष्टि से महत्वपूर्ण है। इसके अलावा, ग्रेंजर कारण-कार्य संबंध परीक्षण के परिणाम भी संकेत देते हैं कि मनोभाव सूचकांक के पास मुद्रास्फीति का अनुमान लगाने की पर्याप्त क्षमता है।

संदर्भ

Baker, S. R., Bloom, N. and Davis, S. J. (2016), "Measuring economic policy uncertainty", *The Quarterly Journal of Economics,* 131(4), 1593-1636.

Baker, S. R., Bloom, N., Davis, S. J. and Kost, K. J. (2019), "Policy News and Stock Market Volatility", *National Bureau of Economic Research (NBER) Working Paper* No. 25720

Beckers, B., Kholodilin, K. A. and Ulbricht, D. (2017), "Reading between the Lines: Using Media to Improve German Inflation Forecasts", *German Institute for Economic Research, Discussion Papers* 1665

Bhagat, S., Ghosh, P. and Rangan, S P. (2013), "Economic Policy Uncertainty and Economic Growth in India", *Indian Institutes of Management (IIM) Bangalore Working Paper* No. 407

Carroll, C. D. (2003), "Macroeconomic Expectations of Households and Professional Forecasters", *The Quarterly Journal of Economics,* 118(1)

Chakraborty, C., and Joseph, A. (2017), "Machine learning at central banks", *Bank of England, Working Paper* No. 674

Ehrmann, M., Pfajfar, D. and Santoro, E. (2017), "Consumers' Attitudes and Their Inflation Expectations", *International Journal of Central Banking*

Hendry, S. (2012): "Central Bank Communication or the Media's Interpretation: What Moves Markets?" *Bank of Canada, Working Paper 9*

Iglesias, J., Ortiz, A. and Rodrigo, T. (2017), "How do the Emerging Markets Central Bank talk? A Big Data approach to the Central Bank of Turkey", *BBVA Economic Research Department Working Paper*, 17-24

Lamla, M. J. and Lein, S. M. (2008), "The Role of Media for Consumers' Inflation Expectation Formation", *Swiss Economic Institute (KOF) Working Paper,* No. 201

Lamla, M. J. and Maag, T. (2012), "The Role of Media for Inflation Forecast Disagreement of Households and Professional Forecasters", *Journal of Money, Credit and Banking,* Vol. 44, No. 7

Lamla, M. J. and Sturm, J. E. (2013), "Interest rate expectations in the media and central bank communication," *KOF Working Paper*, No. 334

Lekshmi, O. and Mall, O.P. (2015), "Forward looking surveys for tracking Indian economy: an evaluation", *Bank for International Settlements (BIS) Irving Fisher Committee (IFC) Bulletin* No. 39

Loughran, T. and McDonald, B. (2011), "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance,* 66, 35-65.

Lucca, D. O. and Trebbi, F. (2009), "Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements", *NBER Working Paper* No. 15367

Manela, A. and Moreira, A. (2017), "News implied volatility and disaster concerns" *Journal of Financial Economics,* 123, 137-162

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P. and Smith, R. (2018), "News and narratives in financial systems: exploiting big data for systemic risk assessment", *Bank of England Working Paper* No. 704

Picault, M. and Renault, T. (2017), "Words are not all created equal: A new measure of ECB communication", *Journal of International Money and Finance* 79, 136–156.

Shapiro, A. H., Moritz, S., and Wilson, D. (2017), "Measuring News Sentiment", *Federal Reserve Bank of San Francisco, Working Paper* 01

Tobback, E. and Nardelli, S. and Martens, D. (2017), "Between Hawks and Doves: Measuring Central Bank Communication", *European Central Bank (ECB) Working Paper* No. 2085

Turney, P. D. (2002), "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417-424).

अनुबंध I – मनोभाव वर्गीकरण

ऑनलाइन प्रिंट मीडिया से एकत्र की गई कच्ची खबरों को यूनिफ़ॉर्म रिसोर्स लोकेटर (यूआरएल), विशेष वर्णों और प्रतीकों, विराम चिह्न और अंकों को हटाकर उपयुक्त रूप से परिष्कृत किया गया है। कुछ कृत्रिम महत्वहीन शब्दों के साथ सामान्य महत्वहीन शब्दों[1], जो मनोभाव वर्गीकरण के लिए कोई विशिष्ट अर्थ नहीं रखते हैं, को भी हटा दिया गया था।

पाठ में शब्दों को व्याकरण के आधार पर पठनीयता के उद्देश्य से अक्सर अलग-अलग रूपों में प्रयोग किया जाता है जिसका अंतर्निहित अर्थ समान होता है। समान अर्थ रखने वाले कई शब्दों से एक महत्वपूर्ण लेख तैयार करने के उद्देश्य से शब्दों को उनके मूल रूप में लिया जाता है। हमने लेम्माटाइजेशन[2] का उपयोग किया है, जो प्रत्येक शब्द को उसके लेम्मा में रूपांतरित करने की एक प्रक्रिया है तथा जो वास्तविक भाषा का शब्द है। रूपांतरण करते समय किसी शब्द के व्याकरण, शब्दावली और शब्दकोश के महत्व का उचित ध्यान रखा जाता है।

समग्र कॉर्पस के लिए, तब लेम्माटाइज़्ड समाचारों का टोकनाइजेशन यूनीग्राम (व्यक्तिगत शब्द) में किया जाता है, जिसके परिणामस्वरूप प्रत्येक दस्तावेज के लिए कुछ शब्दों का निर्माण होता है।

टर्म फ्रीक्वेंसी - इनवर्स डॉक्यूमेंट फ्रीक्वेंसी (टीएफ-आईडीएफ) भार का उपयोग कर शब्दों को भारित किया गया है – जैसा कि नीचे परिभाषित किया गया है:

$$W_{ij} = TF_{ij} \times \log_e\left(\frac{N}{DF_i}\right)$$

जहां TF_{ij} = दस्तावेज़ j में i शब्द कितनी बार दोहरायी गई है

DF_i = i शब्द युक्त दस्तावेजों की संख्या

N = कुल दस्तावेजों की संख्या

टीएफ-आईडीएफ भार एक माप है जिसका दिए गए कॉर्पस में शब्द के महत्व का मूल्यांकन करने के लिए शाब्दिक डेटा में अक्सर उपयोग किया जाता है। किसी शब्द को उच्च भार तब दिया जाता है जब वह दस्तावेज (टीएफ द्वारा) में बार-बार आता है, लेकिन कॉर्पस में दस्तावेजों की संख्या इसे बराबर कर देती है जिसमें वह शब्द (आईडीएफ द्वारा) होता है, परिणामस्वरूप भार में संतुलन आता है।

यद्यपि एसवीएम अरैखिक निर्णय सीमाओं को संभाल सकता है, डेटा के स्वरूप को देखते हुए, हम इस लेख में रैखिक एसवीएम का उपयोग करते हैं। रैखिक एसवीएम वर्गीकरण मॉडल एक अधिकतम मार्जिन क्लासिफायर है और इसके निम्न रूप हैं:

$$f(x_j) = w_0 + w_{1j}x_{1j} + w_{2j}x_{2j} + \cdots + w_{nj}x_{nj}$$

जहां w_{ij} = दस्तावेज j में i शब्द का भार

x_{ij} = दस्तावेज j में i शब्द का आना

w_0 = अवरोधन

परिणामी निर्णय फंक्शन $f(x_j)$ का चिन्ह किसी अमुक दस्तावेज़ का पूर्वानुमानित वर्ग है।

निर्णय फंक्शन का भार केवल प्रशिक्षण डेटा सेट के सबसेट का एक फंक्शन है, जिसे *सपोर्ट वैक्टर* कहा जाता है। वे डेटा बिंदु हैं जो निर्णय सीमा के सबसे करीब हैं और मार्जिन पर होते हैं। मॉडल का प्रशिक्षण करते समय विभिन्न शब्दों का भार प्राप्त किया जाता है। चूंकि हमारे पास चार-वर्गीय वर्गीकरण समस्या है, इसलिए छह एक बनाम एक बाइनरी सब-क्लासिफायर तैयार किए जाते हैं, और अंतिम मनोभाव वर्ग को अधिकतम मतों के आधार पर चुना जाता है।

ट्रेनिंग डेटा को गलत वर्गीकरण से होने वाला नुकसान, नुकसान पैरामीटर सी पर आधारित है। सी का मान अधिक होने से गलत वर्गीकरण से होने वाला नुकसान बढ़ जाता है, जबकि सी का मान कम होने से गलत वर्गीकरण की त्रुटि कम हो जाती है। सी का एक इष्टतम मान होना जरूरी है, जिसे पैरामीटर ट्यूनिंग का उपयोग करके प्राप्त किया जा सकता है।

---

[1] कम महत्वपूर्ण शब्द प्राकृतिक भाषा के शब्द होते हैं, जो पाठ में अक्सर आते हैं, हालांकि पाठ के लिए वे थोड़ा ही अर्थ रखते हैं, उदाहरण के लिए, “the”, “a”, “and”, “this”, “are” आदि। हमने स्मार्ट (‘सिस्टम फॉर दी मैकेनिकल एनालिसिस एंड रिट्रायवल ऑफ टेक्स्ट’) से अंग्रेजी के कम महत्वपूर्ण शब्दों का उपयोग किया जिसमें शाब्दिक विश्लेषण में व्यापक रूप से उपयोग किए जाने वाले कम महत्वपूर्ण शब्दों के सेट शामिल हैं।

[2] लेम्माटाइजेशन के लिए R पैकेज TEXTSTEM का उपयोग किया गया है।

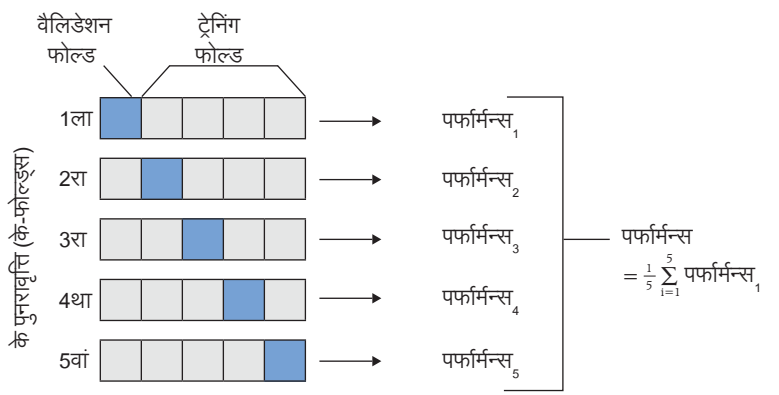
अनुबंध I – मनोभाव वर्गीकरण *(समाप्त)*

वर्तमान मामले में, K = 10 वाले के-फोल्ड क्रॉस-वैलिडेशन का उपयोग करके सी का इष्टतम मान प्राप्त किया गया है, विवरण निम्नानुसार हैं:

के-फोल्ड क्रॉस-वैलिडेशन

- यादृच्छिक ढंग से ट्रैनिंग डेटासेट को के "फोल्ड" में विभाजित करें
- प्रत्येक के-फोल्ड (वैलिडेशन) के लिए, डेटासेट के K-1 फोल्ड पर मॉडल विकसित करें
- इसके उपरांत K^{th} फोल्ड की प्रभावशीलता की जांच करने के लिए मॉडल का परीक्षण करें
- प्रत्येक अनुमान के पर्फार्मन्स (त्रुटि) को रिकॉर्ड करें
- इसे तब तक दोहराएं जब तक प्रत्येक के-फोल्ड्स वैलिडेशन सेट के रूप में कार्य नहीं करता
- के दर्ज त्रुटियों के औसत को क्रॉस-वैलिडेशन त्रुटि कहा जाता है और मॉडल के लिए पर्फार्मन्स मेट्रिक के रूप में कार्य करता है

के फोल्ड क्रॉस वैलिडेशन (के =5) का चित्रण



स्रोत : http://ethen8181.github.io/machine-learning/model_selection/model_selection.html

अनुबंध II - विशेषता चयन पद्धति

(i) दस्तावेज बारंबारता आधारित उपाय - पाठ्य दस्तावेजों में, ऐसे शब्दों का होना एक आम प्रक्रिया है जिनका उपयोग अक्सर सभी दस्तावेजों में नहीं किया जाता है। शब्दों (विशेषताओं) को संबंधित दस्तावेज की बारंबारता (यानी, ऐसे दस्तावेजों का अनुपात जिसमें वह अमुक शब्द आता है) के आधार पर श्रेणीबद्ध किया जा सकता है। थ्रेशोल्ड वैल्यू का उपयोग करके, उन कछ एक शब्दों को बाहर किया जा सकता है जिन्हें निचली श्रेणी में रखा गया है। हमने थ्रेशोल्ड वैल्यू को 0.005 के रूप में उपयोग किया है, जिसका अर्थ है दस्तावेजों में 0.5 प्रतिशत से कम बार आने वाले उन सभी शब्दों को बाहर रखा गया है।

(ii) ची-स्क्वायर उपाय - ची स्क्वायर आंकड़ा χ^2 विशेषता और वर्ग के बीच के संबंध को मापता है। नीचे बताए गए सेल फ्रीक्वेंसी के रूप में समाचारों की गिनती वाले आकस्मिक सारणी का उपयोग करके, χ^2 की गणना की जाती है और पी-मान प्राप्त किया जाता है। पर्याप्त χ^2 मान का तात्पर्य है कि संबंधित वर्ग दिए गए विशेषता से अधिक जुड़ा हुआ है। जो विशेषताएं संबंधित वर्ग (1 प्रतिशत के स्तर पर पर्याप्त) के साथ पर्याप्त रूप से नहीं जुड़ी पाई गईं, उन्हें छोड़ दिया गया है।

	लेबल			
विशेषता	**कमी**	**वृद्धि**	**तटस्थ**	**शून्य**
उपस्थित	X11	X12	X13	X14
अनुपस्थित	X21	X22	X23	X24