International Journal of Hindi Research
ISSN: 2455-2232, Impact Factor:RJIF 5.22
www.hindijournal.com
Volume 2; Issue 5; September 2016; Page No. 07-10

मुहावरेदार अभिव्यक्तियों के भाव–विश्लेषण का साहित्य पुनरावलोकन

(A Literature Review on Sentiment Analysis of Idiomatic Expressions)

प्रितेंद्र कुमार मालाकार

शोधार्थी, प्रौद्योगिकी अध्ययन केंद्र, म.गां.अं.हिं.वि., वर्धा, महाराष्ट्र, भारत।

सारांश

प्रस्तुत शोधपत्र के माध्यम से मुहावरेदार अभिव्यक्तियों के भाव–विश्लेषण से संबंधित पूर्व में किए गए शोधकार्यों का एक संक्षिप्त पुनरावलोकन प्रस्तुत किया गया है जिसमें शोधकार्यों को वैश्विक तथा स्थानीय (भारतीय) भाषाओं के आधार पर वर्गीकृत किया गया है। इस शोधपत्र में मुहावरेदार अभिव्यक्तियों के भाव–विश्लेषण हेतु प्रयुक्त की गई विभिन्न प्रविधियों एवं संसाधनों का अध्ययन एवं विश्लेषण किया गया है।

मूलशब्द: भाव–विश्लेषण, प्राकृतिक भाषा संसाधन, मुहावरेदार अभिव्यक्ति, सोशल मीडिया

प्रस्तावना

विश्व इंटरनेट सांख्यिकी के अनुसार, विश्व की कुल जनसंख्या के लगभग 49.2 % लोग इंटरनेट से जुड़े हुए हैं [19] जिसमे से लगभग 31.2 % लोग सोशल मीडिया का प्रयोग करते हैं [20]। सोशल मीडिया का प्रयोग लोगों के द्वारा अपनी दैनिक जीवन से जुड़ी समस्याओं को हल करने, महत्वपूर्ण निर्णय लेने, विभिन्न राजनीतिक या सामाजिक मुद्दों पर परामर्श लेने/देने, किसी विषय/व्यक्ति/उत्पाद या सेवा के संबंध में प्रतिक्रिया व्यक्त करने, अध्ययन आदि में भी किया जा रहा है[17] जिससे सोशल मीडिया पर व्यक्त की गई प्रतिक्रियाओं की संख्या में वृद्धि हुई है। इंटरनेट पर ये प्रतिक्रियाएं स्थायी तथा डिजिटल रूप से उपलब्ध होती हैं [1, 2, 3]। पिछले कुछ दशकों में प्रतिक्रियाओं की इस विशाल मात्रा ने शोध समुदायों का ध्यान अपनी ओर आकर्षित किया है क्योंकि इनके अध्ययन एवं विश्लेषण से प्राप्त महत्वपूर्ण सूचनाओं का प्रयोग शोध, राजनीतिक, सामाजिक, व्यावसायिक आदि विभिन्न क्षेत्रों में किया जा सकता है [3]। उक्त प्रतिक्रियाओं के संसाधन, विश्लेषण एवं वर्गीकरण हेतु विभिन्न प्रकार की प्राकृतिक भाषा संसाधन आधारित पद्धतियों का प्रयोग किया जाता है जिसमें से भाव–विश्लेषण प्रमुख है।

1. भाव–विश्लेषण

भाव–विश्लेषण एक संगणकीय प्रक्रिया है जिसका प्रयोग किसी पाठ का भाव–वर्गीकरण करने के लिए किया जाता है। भाव–विश्लेषण में प्राकृतिक भाषा संसाधन आधारित संक्रियाओं की सहायता से पाठ में उपस्थित भावयुक्त अभिलक्षणों का संसाधन एवं विश्लेषण किया जाता है तत्पश्चात पाठ को भाव के आधार पर सकारात्मक, नकारात्मक या निष्पक्ष वर्ग में विभाजित किया जाता है।

1.1. पाठ वर्गीकरण

भाव–विश्लेषण में किसी पाठ को निम्न प्रकार से वर्गीकृत किया जाता है:

1.1.1. वस्तुनिष्ठ पाठ–

ऐसे पाठ जिसमें वक्ता के भाव समाहित होते हैं, वस्तुनिष्ठ पाठ कहलाते हैं। वस्तुनिष्ठ पाठ को तीन ध्रुवण स्तर पर विभाजित किया जाता है:

- **सकारात्मक**
  ऐसे पाठ जिसमे वक्ता के सकारात्मक भाव उपस्थित हों, सकारात्मक पाठ कहलाते हैं।
  उदाहरण– तारे ज़मीन पर ***अच्छी*** मूवी है।
- **नकारात्मक**
  ऐसे पाठ जिसमे वक्ता के नकारात्मक भाव उपस्थित हों, नकारात्मक पाठ कहलाते हैं।
  उदाहरण– मोहन ने ***खराब*** गेंदबाजी की।
- **निष्पक्ष**
  ऐसे पाठ जिसमे वक्ता के न तो सकारात्मक और न ही नकारात्मक भाव प्रदर्शित होते हैं, निष्पक्ष पाठ कहलाते हैं।
  उदाहरण– दोपहर के बाद मुझे ***भूख*** लगने लगती है।

1.1.2. तथ्यात्मक पाठ–

ऐसे पाठ जिसमें केवल सामान्य सूचनाएं होती हैं तथा पाठ में वक्ता के किसी भी प्रकार के भाव समाहित नहीं होते हैं, तथ्यात्मक पाठ कहलाते हैं।
उदाहरण– रेल बजट की घोषणा होने वाली है।

2. समस्या कथन

हिंदी भाषा के भाव–विश्लेषण संबंधित अधिकतर शोधकार्य केवल भावयुक्त अभिलक्षणों (भाववाचक संज्ञा, विशेषण, क्रिया–विशेषण, नकारात्मक शब्दों) के विश्लेषण तथा वर्गीकरण पर आधारित हैं किंतु राजनीतिक, सामाजिक या चुनावी मुद्दों पर जनता अपनी टिप्पणियों में मुहावरेदार अभिव्यक्तियों का प्रयोग भी करती है जो कि विषय एवं भाव की दृष्टि से सकारात्मक, नकारात्मक या निष्पक्ष होते हैं[14]।

उदाहरण

1) दोहरे मापदंड वाले दलों (भाजपा–कांग्रेस) को अब समझना होगा। केजरीवाल ने ***चौके–छक्के*** मार दिए तो परेशानी हो जाएगी [18]।
2) महंगाई ने लोगों की ***कमर तोड़*** दी है।
3) रमेश का व्यवहार ***न उधौ से लेना न माधो को देना*** जैसा है।

(उपर्युक्त वाक्यों में रेखांकित मुहावरे क्रमश: सकारात्मक, नकारात्मक एवं निष्पक्ष ध्रुवणता प्रकट करते हैं)।

अत उक्त समस्या के समाधान हेतु ऐसे संगणकीय अभिगम या मॉडल की आवश्यकता है जो पाठ में से मुहावरेदार अभिव्यक्तियों

को स्वत: ही चिन्हित कर उसका भाव–वर्गीकरण (सकारात्मक, नकारात्मक या निष्पक्ष) करे जिससे भाव–विश्लेषण प्रक्रिया से प्राप्त परिणामों की शुद्धता व गुणवत्ता में वृद्धि हो सके।

3. साहित्य पुनरावलोकन

प्रस्तुत शोधकार्य मुख्यत मुहावरेदार अभिव्यक्तियों के भाव–विश्लेषण पर केंद्रित है इसलिए साहित्य पुनरावलोकन हेतु केवल उन्हीं शोधपत्रों, शोधप्रबंधों का चयन किया गया है जिनमें मुहावरेदार अभिव्यक्तियों के भाव–विश्लेषण संबंधी शोधकार्यों का वर्णन मिलता है। इस पुनरावलोकन को विदेशी भाषाओं तथा भारतीय भाषाओं के परिप्रेक्ष्य के अनुसार वर्गीकृत किया गया है।

3.1. विदेशी भाषाओं के परिप्रेक्ष्य में

1. Hossam S- Ibrahim एवं समूह ने मुहावरेदार अभिव्यक्तियों के पहचान एवं वर्गीकरण हेतु भावकोश के अंतर्गत आधुनिक मानक अरबी के बोलचाल में प्रयुक्त मुहावरों एवं लोकोक्तियों का संग्रहण किया। भावकोश में संग्रहित डाटा का वाक्यस्तरीय विश्लेषण करते हुए अर्थ के आधार पर सकारात्मक एवं नकारात्मक दो वर्गों में विभाजन किया गया। उन्होने एन–ग्राम तथा सिमिलरिटी मापन विधि का प्रयोग करते हुए पाठ में से मुहावरों, लोकोक्तियों एवं पदों का प्रत्ययन किया। उक्त प्रस्तावित प्रणाली की शुद्धता औसतन 88.78% आंकी गई [9]।
2. Taysir Hassan A- Soliman एवं समूह ने युवाओं के द्वारा सोशल मीडिया (फेसबुक तथा ट्विटर) पर व्यक्त किए जाने वाले अरबी भाषा के अनौपचारिक अभिव्यक्तियों (जैसे– नवसृजित शब्द, मुहावरे आदि) का भाव–विश्लेषण करने हेतु SVM based classifier विधि प्रस्तावित की [10]।

   उक्त विधि निम्न तीन चरणों में क्रियान्वित की गई–
   1) डाटा निर्माण
   2) डाटा पूर्व–संसाधन
   3) डाटा वर्गीकरण

   उपरोक्त विधि के अतिरिक्त नवसृजित शब्दों तथा मुहावरेदार अभिव्यक्तियों का एक लेक्सीकॉन SSWIL (Slang Sentimental Words and Idioms Lexicon) बनाया गया। उक्त प्रविधि (SSWIL with Classic Classification) को यादृच्छिक 150 प्रतिक्रियाओ पर क्रियान्वित करने से प्रणाली का Precision, Recall, F-Measure] Specificity क्रमश 88.63%, 78%, 82. 97%, 54.54%मापा गया।
3. Vassiliki Rentoumi के द्वारा [12] अंग्रेजी भाषा के Metaphorical Expressions को नियंत्रित करने हेतु SentiFig नामक मशीन लर्निंग विधि का प्रयोग किया गया है जो निम्न तीन चरणों में क्रियान्वित होता है।
   1) शब्द–भावों का विसंदिग्धीकरण (WSD)
   2) शब्द–भावों को ध्रुवणता प्रदान करना (SLP)
   3) वाक्यस्तरीय ध्रुवणता की पहचान करना
4. अक्षत बक्लीवाल एवं समूह के द्वारा सोशल मीडिया पर दी जाने वाली राजनीतिक प्रतिक्रियाओं में प्रयुक्त किए जाने वाले अंग्रेजी भाषा के कुल 89 पदो (भावयुक्त मुहावरेदार अभिव्यक्तियों) को भाव–विश्लेषण में सम्मिलित किया तथा 58% शुद्धता प्राप्त की [11]।
5. टिम वान एवं समूह के द्वारा वृहद कार्पोरा से बड़े पैमाने पर मुहावरेदार अभिव्यक्तियों के प्रत्ययन हेतु एक अनिर्देशित एवं स्वचालित विधि प्रस्तावित किया गया [7]।
6. यूलिया तथा शूली के द्वारा लघु समानांतर कार्पोरा से बहुशब्दीय अभिव्यक्तियों के विभिन्न स्वरूपों जैसे कि– मुहावरे, लोकोक्ति, पुनरुक्ति आदि के प्रत्ययन एवं लक्ष्य भाषा में अनुवाद हेतु एक सामान्य प्रविधि प्रस्तुत की। उनके द्वारा लघु द्विभाषीय कार्पस लिया गया तथा दोषपूर्ण शब्द अलाइनमेंट में से बहुशब्दीय अभिव्यक्तियों का प्रत्ययन किया। इसके पश्चात एक वृहद एकभाषीय कार्पस से बहुशब्दीय अभिव्यक्तियों का चयन करने एवं वरीयता प्रदान करने हेतु सांख्यिकीय विधि का प्रयोग किया [12]।
7. Lowri Williams एवं समूह [14] ने भाव–विश्लेषण के स्वचालित अभिगमों में मुहावरेदार अभिव्यक्तियों की भूमिका का अध्ययन किया एवं पाया कि मुहावरेदार अभिव्यक्तियों का भाव–वर्गीकरण करने से पारंपरिक भाव–विश्लेषण के परिणाम में सुधार आ सकता है। उन्होने परिणामों की तुलना निम्न दो विधियों से की –
   1) भाव–विश्लेषण से संबंधित 580 मुहावरों को संकलित करने हुए उनका प्रयोग एक फीचर के प में किया एवं प्रत्येक मुहावरे की भाव के प में मैपिंग की गई। इन मैपिंग को प्राप्त करने हेतु एक वेब आधारित अभिगम का प्रयोग किया गया तत्पश्चात पांच स्वतंत्र एनोटेटर्स ने Krippendorff 's गुणांक का प्रयोग करते हुए crowdsourcing सूचना के गुणवत्ता का आंकलन किया।
   2) भाव–विश्लेषण के परिणाम का मूल्यांकन करने हेतु ऐसे वाक्यों का कार्पस निर्मित किया जिनमें मुहावरेदार अभिव्यक्तियों का प्रयोग हुआ हो तथा प्रत्येक वाक्य को भाव के रूप में एनोटेट किया। उपर्युक्त प्रविधि का प्रिसिजन तथा रिकाल क्रमश 64% एवं 61% प्राप्त किया गया।
8. Kong & Joo Lee एवं समूह [16] ने कोरियन भाषा के बहुशब्दीय भाव अभिव्यक्तियों के स्वत उत्पादन हेतु एक मूल भाव कोश (Seed Sentiment Lexicon) तथा एक वृहद–मात्रात्मक क्षेत्र–विशेष कार्पस का प्रयोग किया है। उक्त भावकोश में केवल एकल भाव शब्दों को संग्रहित किया गया। उनकी परिकल्पना के अनुसार बहुशब्दीय अभिव्यक्तियां एकल शब्दों के विस्तार से प्राप्त होती हैं जिनमें वे शब्द स्वयं तथा उनके क्रमागत शब्द (जिनका प्रयोग बहुशब्दीय अभिव्यक्तियों के ध्रुवणता निर्धारण में किया जा सकता है) भी सम्मिलित होते हैं। उनके उन्होने एक अनिर्देशित (Unsupervised) मॉडल का प्रयोग किया जो कि निम्न पांच चरणों में क्रियान्वित होता है–
   1) पूर्व–संसाधन
   2) आधारभूत भाव–विश्लेषण
   3) भाव अभिव्यक्तियों का विस्तार
   4) सामान्यीकरण एवं चयन
   5) Decision of Lexicon Entries

   प्रणाली के पांचवें चरण में यह निर्धारित किया जाता है कि प्राप्त बहुशब्दीय अभिव्यक्तियों में से किन–किन अभिव्यक्तियों को कोश में स्थान दिया जाए। उक्त प्रविधि का प्रिसिजन तथा रिकाल क्रमश 61.5% तथा 58.1% प्राप्त किया गया।
9. Antonio Moreno-Ortiz एवं समूह [17] के द्वारा स्पेनिश भाषा के बहुशब्दीय अभिव्यक्तियों के भाव–विश्लेषण हेतु 'Sentitext' नामक कोश आधारित अनुप्रयोग का निर्माण किया गया। इस वेब आधारित अनुप्रयोग को सी++ तथा पाईथान प्रोग्रामिंग भाषा की सहायता से विकसित किया था।

इस अनुप्रयोग के अंतर्गत निम्न तीन प्रमुख संसाधनों का प्रयोग किया गया था–

1) Individual Word Dictionary
2) Multiword Dictionary
3) Context Ruleset

3.2. भारतीय भाषाओं के परिप्रेक्ष्य में

हिंदी भाषा के भाव–विश्लेषण संबंधी शोधकार्य अन्य भाषाओं (विशेषकर अंग्रेजी) की अपेक्षाकृत कम हुए हैं। अभी तक हिंदी पाठ के भाव–विश्लेषण के क्षेत्र में किए गए शोधकार्यों में से कुछ प्रमु ख निम्न हैं–

1. मोनिका गाले एवं समूह के द्वारा मुहावरेदार अभिव्यक्तियों की पहचान हेतु एक नियम आधारित अभिगम प्रस्तावित किया गया तथा मुहावरेदार अभिव्यक्तियों के अनुवाद हेतु गूगल अनुवाद प्रणाली का प्रयोग किया गया। उक्त प्रणाली की शुद्धता 70: आंकी गई [4]।
2. मोनिका शर्मा एवं विशाल गोयल के द्वारा हिंदी से पंजाबी मशीनी अनुवाद में लोकोक्तियों (Proverbs) के प्रत्ययन हेतु एक ग्राफिकल यूजर इंटरफेस का निर्माण किया गया। उन्होने रिलेशनल डाटा अभिगम का प्रयोग किया। हिंदी तथा पंजाबी के लोकोक्तियों को स्थायी एवं परिवर्तनीय (Dynamic) के रूप में विभाजित किया। स्थायी वर्ग को रेगुलर एक्सप्रेशंस के द्वारा नियंत्रित किया गया तथा परिवर्तनीय वर्ग के अंतर्गत लोकोक्तियों के सभी संभव रूपों को रखा गया। स्थायी वर्ग के लोकोक्तियों का डाटाबेस से मिलान किया गया; डाटाबेस से मिलान होने पर लोकोक्तियों का पंजाबी अर्थ प्रदर्शित किया गया। इस अभिगम से प्राप्त परिणाम की शुद्धता लगभग 60–80% मापी गई [5]।
3. अश्वनी अग्रवाल एवं समूह के द्वारा बंगाली भाषा के कार्पस से मुहावरेदार अभिव्यक्तियों के स्वत प्रत्ययन हेतु रूपवैज्ञानिक विश्लेषण एवं सांख्यिकीय आधारित एक अभिगम प्रस्तावित किया गया जिसमें सभी संज्ञा+क्रिया, विशेषण+क्रिया, क्रिया–विशेषण+क्रिया समूहों तथा मुहावरेदार अभिव्यक्तियों को चिन्हांकित कर प्रत्येक को सांख्यिकीय पैरामीटर जैसे कि– सह–पुनरावृत्ति अथवा आवृत्ति के आधार पर विशिष्ट आंकिक मान प्रदान किया गया [6]।
4. विवेक दुबे एवं समूह ने अंग्रेजी तथा हिंदी भाषा में बहुशब्दीय अभिव्यक्तियों के प्रभाव को समझने के लिए बहुशब्दीय अभिव्यक्तियों की संरचना का अध्ययन एवं विश्लेषण किया। उन्होने बहुशब्दीय अभिव्यक्तियों के संगणन हेतु सांख्यिकीय, भाषावैज्ञानिक, संकरित तथा मशीन लर्निंग विधियों को प्रस्तावित किया [13]।

4. निष्कर्ष

प्रस्तुत शोधपत्र में मुहावरेदार अभिव्यक्तियों के विश्लेषण एवं वर्गीकरण से संबंधित पूर्व में किए गए शोधकार्यों तथा उनमें प्रयुक्त विभिन्न विधियों, तकनीकों तथा अभिगमों का संक्षिप्त विवरण प्रस्तुत किया गया है जिससे हिंदी भाषा के मुहावरेदार अभिव्यक्तियों के भाव–विश्लेषण हेतु उचित एवं प्रभावी प्रविधि निर्माण में सहायता मिलेगी।

5. संदर्भ सूची

1. Arora, Piyush. Sentiment Analysis for Hindi Language (MS Thesis). IIIT Hyderabad, 2013.
2. Joshi A, Balamuraly AR, Pushpak Bhattacharya. A Fallback Strategy for Sentiment Analysis in Hindi a Case Study. Proceedings of ICON 8th International Conference on Natural Language Processing Macmillan Publishers India, 2010.
3. Pang B, Lillian Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2008; 2(1-2)1-135.
4. Gaule, Monika, Dr. Gurpreet Singh Josan. Machine Translation of Idioms from English to Hindi. International Journal of Computational Engineering Research. 2012, 2-6.
5. Sharma, Monika, Vishal Goyal. Extracting Proverbs in Machine translation from Hindi to Punjabi using Relational Data Approach. International Journal of Computer Science and Communication.2011; 2611-613.
6. Agarwal, Aswhini, Biswajit Ray. Automatic Extraction of Multiword Expressions in Bengali An Approach for Miserly Resource Scenarios. Proceedings of the International Conference on Natural Language Processing. (ICON 2004). Allied Publishers. 2004, 165-172.
7. Moiron, Begona Villada, Tim Van de Cruys. Semantics-based Multiword Expression Extraction. Proceedings of the Workshop on a Broader Perspectives on Multiword Expressions. Associations for Computational Linguistics. 25-32.
8. Taysir Hassan A, Soliman. *et al*. Mining social networksarabic slang comments. In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM 13), Prague, Czech Republic.
9. [Ibrahim] Hossam S, Sherif M, Abdou, Mervat Gheith. Idioms&Proverbs Lexicon for Modern Standard Arabic and Colloquial Sentiment Analysis- International Journal of Computer Applications. 2015; (0975 -8887)118-11.
10. Rentoumi, Vassiliki. Sentiment Analysis of Metaphorical Language- PhD Thesis- University of the Aegean-
11. Bakliwal, Akshat, Jennifer Foster, Jennifer van der Puil, Ron O Brien, Lamia Tounsi, Mark Hughes. Sentiment Analysis of Political Tweets Towards an Accurate Classifier. Proceedings of the Workshop on Language in Social Media (LASM 2013), 2013, 49-58.
12. Tsvetkov, Yulia, Shuly Wintner. Extraction of Multi&word Expressions from Small Parallel Corpora. 2010 Coling 2010, 256-1264. Beijing.
13. Dubey, Vivek, Pankaj Raghuwanshi, Sapna Vyas- Impact of Multiword Expression in English&Hindi Language. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2015; 4(3)101-105.
14. Williams, Lowri, Christian Bannister A, Michael Arribas&Ayllon b, Alun Preece a] Irena Spasic. The role of idioms in sentiment analysis- Expert Systems with Applications. 2015; 427375-7385.
15. Lee, Kong-Joo, Jee-Eun Kim, Bo-Hyun Yun. Extracting Multiword Sentiment Expressions by Using a Domain&Specific Corpus and a Seed Lexicon- ETRI Journal. 2013; 35(5)838-848.
16. Moreno-Ortizl, Antonio, Chantal Pérez-Hernández, Ángeles Del-Olmo M. Managing Multiword Expressions in a Lexicon&Based Sentiment Analysis System for Spanis- Proceedings of the 9th Workshop on Multiword Expressions. MWE 2013, 1-10.
17. डॉ. कुमुदिनी पति. सोशल नेटवर्किंग है नया औजार. समीरा (पत्रिका), CHHHIN\2009\27629, अंक 5, मई 2014, पृष्ठ क्रं 07–08.

18. दैनिक भास्कर (ई–पेपर). Weblink www-epaper-bhaskar-com- Visited on 24 December, 2013.
19. विश्व इंटरनेट सांख्यिकी. Weblink. http@@www-internetworldstats-com@stats- htm- Visited on 07 Septmber] 2016.
20. http／／www-worldometers-info@world&population@Visited on 07 Septmber, 2016.