भारतीय वैज्ञानिक एवं औद्योगिक अनुसंधान पत्रिका
वर्ष 29 अंक (1) जून 2021 पृ. 43-48

हिन्दी में तंज़ का मशीन लर्निंग तकनीकों द्वारा प्रदर्शन-मूल्यांकन

प्रज्ञा कात्यायन एवं निशीथ जोशी
कम्प्यूटर साइंस विभाग, वनस्थली विद्यापीठ 304 022 (राजस्थान)

सारांश : मनुष्य अपने मन की बात व्यक्त करने के लिए जिन भावनाओं का सहारा लेते हैं, उन पर तंज़ का खासा प्रभाव देखने को मिलता है। कटाक्ष के माध्यम से अपनी बात सामने रखना आज के समय में बहुत ही प्रचलित है और लोग इसे बहुत ही भारी मात्रा में काम में लेते हैं। भावनात्मक विश्लेषण, जिसे हम राय खनन के नाम से भी जानते हैं, ने तंज़ को हमेशा ही एक चुनौती की तरह अपने सामने खड़ा पाया है। तंज़, जिसे कि समझना अक्सर मनुष्यों के बस की बात भी नहीं होती, भाषा-संस्करण के लिए बहुत ही विकट संकट की भांति प्रतीत होता है। इसके प्रमुख कारणों में से एक है इसकी विविधता। तंज़ प्रस्तुत करने के विविध तरीकों के कारण इसे समझ पाना बेहद कठिन होता है। चूंकि तंज़ रूपी समस्या का कोई बेहतर समाधान अब तक सामने नहीं आया है, इसलिए इस लेख द्वारा प्रख्यात मशीन लर्निंग तकनीकों द्वारा तंज़ के विभिन्न प्रकारों को समझने की प्रक्रिया का मूल्यांकन करने की कोशिश की गयी है।

Performance evaluation of machine learning algorithms for detecting Hindi sarcasm

Pragya Katyayan & Nisheeth Joshi
Faculty of Mathematics & Computing, Banasthali Vidyapith 304 022 (Rajasthan)

Abstract

Sentiments, the common way people express their feelings, have been greatly influenced with the advent of Sarcasm. Being sarcastic is considered trendy and thus people use it extensively in their day-to-day language. Sentiment Analysis, also known as Opinion Mining, has encountered Sarcasm as a challenge since a long time. Sarcasm, which finds few human brains susceptible to its presence and effects, has posed to be the toughest of all problems. One of the issues with Sarcasm Detection is the numerous ways it can be expressed with. Since there has not been a perfect answer to all the Sarcasm issues, this paper attempts to analyse and evaluate the popular Machine learning techniques on mixed sarcasm types.

प्रस्तावना

मनुष्य अपनी विचार भावनाओं के माध्यम से व्यक्त करते हैं। हर व्यक्ति का भावनाएं व्यक्त करने का तरीका एक-दूसरे से अलग होता है। उदाहरण के तौर पर देख जाए तो खुशी की भावना अलग-अलग लोगों द्वारा अलग-अलग तरीकों से व्यक्त की जा सकती है। कोई खुशी के मारे नाचने लगता है तो कोई खुशी के मारे रो देता है। यह पूर्णतया उनके व्यक्तित्व पर निर्भर करता है कि कौन किस प्रकार अपनी भावनाओं को दुनिया के सामने प्रस्तुत करता है। इसी प्रकार लिखित तौर पर भावनाएँ व्यक्त करना भी लोगों के व्यक्तित्व पर निर्भर करता है। भावनाओं के विश्लेषण द्वारा लोगों के लिखे हुए लेखों (ब्लॉग), संक्षिप्त लेखों (माइक्रो-ब्लॉग जैसे ट्विटर), ऑनलाइन शॉपिंग वेबसाइट पर उत्पाद-उल्लेख अथवा किताबों के उल्लेख में इन्हीं छिपी हुई भावनाओं को समझने का प्रयत्न किया जाता है। नवीन सूचना प्रौद्योगिकी ने लोगों को अपने मन की बातें खुल कर कह पाने की आज़ादी दी है और लोगों ने इसका पूरी तरह से फायदा भी

उठाया है। बदकिस्मती से मनुष्य की भावनाओं को सही तरह से जाँचने में कुछ चुनौतियों का सामना करना पड़ता है, जिसमें सबसे बड़ी चुनौती तंज़ है।

तंज़ को हम सकारात्मक और नकारात्मक भावनाओं के बीच की महीन रेखा के रूप में समझ सकते हैं। तंज़ एक ऐसी भाषा-शैली है जिसमें बोलने वाला एक रहस्यमयी तरीके से अपनी भावनाएँ व्यक्त करता है। इसे अस्पष्ट प्रवृत्ति का कहा जाता है और इसी वजह से तंज़ अक्सर समझ से परे होता है। तंज़ का इस्तेमाल अक्सर सुनने वाले व्यक्ति के विरुद्ध टिप्पणी करने के लिए इस्तेमाल किया जाता है, ऐसे शब्दों का चयन करके जो दिखे कुछ और पर उनका अर्थ कुछ और ही समझ आए और साथ ही अन्य स्रोताओं के लिए हास्यपद हो। तंज़ अपने आप में एक गुत्थी है। मौखिक तंज़ लिखित तंज़ से ज़्यादा आसानी से पहचाना जा सकता है। अक्सर देखा गया है कि मनुष्य भी तंज़ को समझ पाने में खुद को असमर्थ पाते हैं, पर जब इसे चेहरे के भावों और बोलने के तरीके के साथ देखा जाता है तो मौखिक तंज़ को पहचान पाना कुछ हद तक आसान हो जाता है। जबकि लिखित तंज़ के साथ अधूरे संदर्भ की वजह से यह समस्या बनी रहती है। तंज़ अक्सर सकारात्मक शब्दों के माध्यम से नकारात्मक भावों को प्रदर्शित करने के लिए काम में लिया जाता है। उदाहरण के लिए 'बेहतरीन चलचित्र!' यह वाक्य देखने-समझने में पूर्ण रूप से सकारात्मक लगता है, परंतु अगर हम इसके संदर्भ के बारे में सोचें तो इस सामान्य से वाक्य के कई अर्थ हो सकते हैं। अगर इस वाक्य को यदि प्रचालित चलचित्र के संदर्भ में कहा गया है, तो यह वाक्य सकारात्मक है, किन्तु अगर यह वाक्य असफल चलचित्र के बारे में कहा गया है, तो यह वाक्य ज़ाहिर तौर पर एक तंज़ है। इस उदाहरण को देख कर हमें यह समझ आता है कि एक वाक्य को सुनकर या पढ़कर हम यह पता नहीं लगा सकते कि वह सकारात्मक है या नकारात्मक भावना को सँजोया हुआ तंज़ है। हालांकि, तंज़ की स्थिति में हर बार वाक्य के बारे में पूरी जानकारी होना भी काफी नहीं होता।

तंज़ का इस्तेमाल करने के तरीके

तंज को इस्तेमाल करने के तीन तरीके हैं:

- चालाकी दिखने हेतु
- अपमानित करने हेतु
- सच को छुपाने हेतु।

इस लेख द्वारा हमने अपमानित करने हेतु इस्तेमाल किए गए तंज़ को पहचानने की कोशिश की है। ऐसे वाक्यों का उद्देश्य सामने वाले को नीचा दिखाने का और उसका मज़ाक उड़ाने का होता है। इस लेख का लक्ष्य अपमानजनक तंज़ वाक्यों का प्रचलित मशीन लर्निंग तकनीकों द्वारा पहचानने का है। हमने सामान्य वाक्यों और तंज़ के वर्गीकरण के लिए नेव बेज़, सपोर्ट वेक्टर मशीन, डिसीजन ट्री तथा न्यूरल नेटवर्क नामक चार प्रचलित वर्गीकारकों का इस्तेमाल किया है। इन सभी के नतीजों का मूल्यांकन करने के लिए प्रिसीजन, रिकॉल तथा एफ़-मेज़र की गणना की जाएगी।

तंज़ का परिचय

हिन्दी में तंज़ का बहुत खास महत्व है। हमेशा से ही मशहूर कवियों और लेखकों ने इसका इस्तेमाल अपने काव्यों, लेखों और कहानियों में किया है और इसके माध्यम से समाज को, हौले से, उसकी कड़वी सच्चाई से अवगत कराया है। तंज़ की सबसे खास बात यह होती है कि वह हमेशा अस्पष्ट रूप में आता है। अक्सर हम इसे सकारात्मक शब्दों के पीछे असली अर्थ छुपाते हुए देख सकते हैं, जिसे सिर्फ वही व्यक्ति समझ सकता है जिसके लिए वह वाक्य कहा गया है। हिन्दी हो या अंग्रेज़ी, तंज़ अपने इसी दोहरे व्यक्तित्व के लिए जाना जाता है और भावनात्मक विश्लेषण और भाषा प्रसंस्करण जैसी विधाओं के लिए एक चुनौती सिद्ध होता है।

कविताओं और कहानियों से निकलकर तंज़ आजकल आम बोल-चाल की भाषा का अभिन्न अंग बन चुका है। लोगों को हर जगह इसका इस्तेमाल करते हुए देखा जा सकता है। इस कारणवश, सटीक भावनात्मक विश्लेषण के लिए तंज़ को सही तरीके से पहचानना बहुत ज़रूरी है।

संबन्धित कार्य

जोशी एवं अन्य (2010) ने हिन्दी में तंज़ का पता लगाने के लिए हिन्दी तंज़ वाक्यों का कोश तयार किया और अंग्रेज़ी के वर्डनेट पर आधारित हिन्दी सेंटी-वर्डनेट बनाया[1]। उन्होंने अलग-अलग दृष्टिकोण के मापदंडों में से सबसे अच्छे मापदण्डों को चुना और उनकी तुलना अंग्रेज़ी के भाव-विश्लेषण मापदण्डों से की है। उनके 3-डी दृष्टिकोण कुछ इस प्रकार हैं:

- **स्व-भाषाई भावना विश्लेषण :** इस दृष्टिकोण में उन्होंने हिन्दी ट्रेनिंग डाटा का इस्तेमाल करके वर्गीकरण प्रारूप (क्लासिफायर) बनाया है।

- **मशीन-अनुवाद भावना विश्लेषण :** इसमें शोधकर्ताओं ने प्रारूप को अंग्रेज़ी कोश पर ट्रेन किया और फिर हिन्दी दस्तावेज़ों को अंग्रेज़ी में अनुवाद करके उस प्रारूप का इस्तेमाल किया।

• **संसाधन-आधारित भावना विश्लेषण :** इस दृष्टिकोण में शोधकर्ताओं द्वारा सेंटी-वर्डनेट के लिए बहुमत पर आधारित वर्गीकरण किया गया है। इस प्रक्रिया में हिन्दी का बहुत बड़ा कोश और मशीन लर्निंग का इस्तेमाल किया गया है।

देसाई एवं दवे (2016) ने हिन्दी तंज़ की खोज में कई चुनौतियों का सामना किया जैसे कि तंज़ डेटासेट की अनुपलब्धता, हिन्दी भाषा के फ्री-ऑर्डर होने के कारण उत्पन्न हुई समस्याएँ, हिन्दी के शब्दों के विभिन्न गुणों के कारण सामने आयी चुनौतियाँ जैसे- समान अर्थ वाले शब्दों का एक से ज़्यादा वर्तनी में उपलब्ध होना, साधन-संसाधन की कमी और सटीक कोश की अनुपलब्धता। उनके दृष्टिकोण में तंज़ को दो वर्गों द्वारा पहचानने की कोशिश की गयी है-

• **टाइप-1 :** तंज़ बताने वाले गुणों की उपस्थिति (इमोजी, हैश टैग, विराम चिन्ह)

• **टाइप-2 :** तंज़ बताने वाले गुणों की अनुपस्थिति।

इन दो तरीकों के लिए शोधकर्ताओं के पास दो प्रकार के कोश थेः

• **टाइप-1 :** 1400 वाक्य।

• **टाइप-2 :** 250 वाक्य।

इनकी विशेषताएँ (फीचर्स) :

• **टाइप-1 :** शाब्दिक, व्यावहारिक, भाषिक विशेषताएँ तथा टी. एफ.-आई. डी. एफ़.

• **टाइप-2 :** विलोम शब्द, सकारात्मक-नकारात्मक शब्दों की सूची, संकेत शब्दों की सूची

निष्कर्ष

• **टाइप-1 :** 5 वर्ग-वर्गीकरण लिब-एस.वी.एम. का इस्तेमाल करके, जिसकी एक्यूरेसी 84% पायी गयी।

• **टाइप-2 :** 5 वर्ग-वर्गीकरण एस.वी.एम. का इस्तेमाल करके, जिसकी एक्यूरेसी 60% पायी गयी।

भारती एवं अन्य (2016) ने हिन्दी तंज़ की खोज के लिए एक नयी तकनीक का प्रस्ताव दिया है जिसमें एक ट्वीट और उससे जुड़ी खबर के बीच के अंतर्विरोध को काम में लिया गया है। इस तकनीक ने समकालीन हिन्दी खबरों और उनसे जुड़े ट्वीट्स पर एक साथ विचार किया है। उन्होंने ट्वीट के संदर्भ के रूप में उससे जुड़ी समकालीन खबर को लिया है और उस तर्ज पर हिन्दी तंज़ को पहचानने की कोशिश की है।

प्रस्तावित कार्य

इस लेख ने विभिन्न मशीन लर्निंग तकनीकों के प्रदर्शन का मूल्यांकन किया है। यह प्रयोग कार्य कोश-संग्रह, डेटा-प्री-प्रोसेसिंग, विशेषता-निष्कर्षण, क्लासिफिकेशन (वर्गीकरण) तथा मूल्यांकन का कार्य करता है।

कोश संग्रह

इस प्रयोग कार्य में 1000 वाक्यों का इस्तेमाल किया गया है जिसमें से 500 तंज़ और बाकी 500 साधारण वाक्य हैं। इन वाक्यों को सोशल मीडिया की विभिन्न साइट्स से लिया गया है, जैसे फेसबुक, ट्विटर और इंस्टाग्राम। इन साइट्स पर ऐसे पृष्ठ उपलब्ध हैं जिन पर खास तौर से सिर्फ तंज़ वाक्य ही डाले जाते हैं, कोश बनाने में मददगार साबित होते हैं। ये सत्यापित वाक्य हैं, इस कारण विश्वसनीय माने जाते हैं। बाकी के 500 वाक्य, जो कि साधारण वाक्य हैं, उन्हें बराबर सकारात्मक तथा नकारात्मक श्रेणी के वाक्यों से बनाया गया है। इस प्रकार बाँटने से हमारे मशीन लर्निंग मॉडल हर तरह के गैर-व्यंग्यात्मक वाक्यों पर खुद को ट्रेन कर सकते हैं और इनके परिणामों में पक्षपात की समस्या से बचा जा सकता है।

वाक्यों को संगृहित करते हुए सबसे बड़ी चुनौती थी इन वाक्यों का टेक्स्ट रूप में न मिलना। हमें बहुत खोजने के बाद भी प्रयोग कार्य के अनुरूप वाक्य नहीं मिल सके क्यूंकि हिन्दी भाषा में तंज़ पर ज़्यादा शोध कार्य नहीं पाये गए हैं। इस प्रकार के वाक्यों को इकट्ठा करना अपने आप में चुनौती भरा कार्य है। हमने सोशल मीडिया साइट्स से जो वाक्य इकट्ठे किए वे अंग्रेज़ी में थे तथा चित्र रूप में मिले। उन्हें प्रयोग कार्य के अनुरूप बनाने के लिए हमने उनका हिन्दी में अनुवाद किया और टेक्स्ट रूप में कोश बनाया। कुछ ऐसा ही सकारात्मक वाक्यों के साथ भी हुआ जहाँ हिन्दी में वाक्य न मिलने की वजह से हमें अंग्रेज़ी के वाक्य ढूंढकर उन्हें हिन्दी में अनुवाद करना पड़ा। जबकि हिन्दी के नकारात्मक वाक्य ऑनलाइन आसानी से उपलब्ध थे, उन्हें बस थोड़े सुधार कार्य की आवश्यकता थी। उनमें से 250 नकारात्मक वाक्यों को हमने अपने कोश में शामिल किया और प्रयोग कार्य के लिए इस्तेमाल किया। इस बात का विशेष ध्यान रखा गया है कि इन सभी वाक्यों की अनुवाद के दौरान प्रकृति न बदले। इनका अर्थ उसी प्रकार से मिलना चाहिए जैसा कि इनके अंग्रेज़ी रूप द्वारा प्राप्त हुए थे।

500 तंज़ वाक्यों को हमने '1' लेबल से अंकित किया है तथा 500 सामान्य वाक्यों को '0' लेबल से अंकित किया है (सारणी 1)।

सारणी 1 – शब्दकोश में वाक्यों को दिये गए लेबल

स्वभाव	क्लास	लेबल
तंज (500)	तंज़	1
सकारात्मक (250)	सामान्य	0
नकारात्मक (250)	सामान्य	0

डेटा प्री-प्रोसेसिंग

टोकेनाइज़ेशन : टेक्स्ट की लंबी कड़ियों को तोड़कर अर्थपूर्ण इकाइयों में बदलने को 'टोकेनाइज़ेशन' कहते हैं।

वाक्य : 'मैं बेवकूफ हुआ करता था लेकिन अब हमारा तलाक हो गया।'

टोकेनाइज्ड

['मैं','बेवकूफ 'ए' हुआ 'ए' करता 'ए' था 'ए' लेकिन 'ए' अब 'ए' हमारा 'ए' तलाक 'ए' हो 'ए' गया 'ए',।']

स्टॉप वड्र्स हटाने की प्रकिया : ऐसे शब्दों को हटाने की प्रकिया जिनका वाक्य के अर्थ में कोई खास योगदान नहीं होता।

हिन्दी के स्टॉप वड्र्स की सूची

['अपना', 'अपनी', 'अपने', 'अभी', 'अंदर', 'आदि', 'आप', 'इत्यादि', 'इन', 'इनका', 'इन्हीं', 'इन्हें', 'इन्हों', 'इस', 'इसका', 'इसकी', 'इसके', 'इसमें', 'इसी', 'इसे', 'उन', 'उनका', 'उनकी', 'उनके', 'उनको', 'उन्हीं', 'उन्हें', 'उन्हों', 'उस', 'उसके', 'उसी', 'उसे', 'एक', 'एवं', 'एस', 'ऐसे', 'और', 'कई', 'कर', 'करता', 'करते', 'करना', 'करने', 'करें', 'कहते', 'कहा', 'का', 'काफ़ी', 'कि', 'कितना', 'किन्हें', 'किन्हों', 'किया', 'किर', 'किस', 'किसी', 'किसे', 'की', 'कुछ', 'कुल', 'के', 'को', 'कोई', 'कौन', 'कौनसा', 'गया', 'घर', 'जब', 'जहाँ', 'जा', 'जितना', 'जिन', 'जिन्हें', 'जिन्हों', 'जिस', 'जिसे', 'जीधर', 'जैसा', 'जैसे', 'जो', 'तक', 'तब', 'तरह', 'तो', 'था', 'थी', 'थे', 'दबारा', 'दिया', 'दुसरा', 'दूसरे', 'दो', 'द्वारा', 'न', 'के', 'नहीं', 'ना', 'निहायत', 'नीचे', 'ने', 'पर', 'पहले', 'पूरा', 'पे', 'फिर', 'बनी', 'बही', 'बहुत', 'बाद', 'बाला', 'बिलकुल', 'भी', 'भीतर', 'मगर', 'मानो', 'मे', 'में', 'यदि', 'यह', 'यहाँ', 'यही', 'या', 'यिह', 'ये', 'रखें', 'रहा', 'रहे', 'लिए', 'लिये', 'लेकिन', 'व', 'वग़ैरह', 'वर्ग', 'वह', 'वहाँ', 'वहीं', 'वाले', 'वुह', 'वे', 'सकता', 'सकते', 'सबसे', 'सभी', 'साथ', 'साबुत', 'साभ', 'सारा', 'से', 'सो', 'संग', 'ही', 'हुआ', 'हुई', 'हुए', 'है', 'हैं', 'हो', 'होता', 'होती', 'होते', 'होना', 'होने', ...]

वाक्य : मैं बेवकूफ हुआ करता था लेकिन अब हमारा तलाक हो गया।

स्टॉप वड्र्स हटाने के बादः ['मैं', 'बेवकूफ', 'अब', 'हमारा', 'तलाक', '।']

स्टेमिंग एवं लेमाटाइज़ेशन : ये दोनों ही शब्दों के मूल रूप से उत्पन्न शब्दों को वापस उनके मूल रूप में बदलने की प्रकियाएँ हैं। इनके बीच एक फर्क है- स्टेम्मिंग से किसी शब्द से जुड़े उपसर्ग अथवा प्रत्यय हटा दिये जाते हैं जिसकी वजह से अक्सर शब्द अर्थ-विहीन रह जाते हैं, किन्तु लेमाटाइज़ेशन से उपसर्ग/प्रत्यय हटाने के बाद उपयुक्त मात्राएँ अथवा अक्षर जोड़ कर उन्हें अर्थ-पूर्ण मूल रूप में बदल दिया जाता है।

उदाहरण

वाक्य : नशेड़ी लोग, बच्चे और पजामी हमेशा सच्चाई बताते हैं।

स्टॉप वड्र्स हटाने के बाद : नशेड़ी लोग बच्चे पजामी हमेशा सच्चाई बताते।

स्टेमिंग : नशेड़ लोग, बच्च पजाम हमेश सच्च बत।

लेमाटाईज़ेशन : नशा लोग, बच्चा पजामी हमेशा सच बात।

लेमाटाइज़ेशन के निष्कर्ष स्टेमिंग की अपेक्षा अच्छे होने के कारण हमने अपने प्रयोग कार्य में केवल इसी तकनीक का इस्तेमाल किया है।

फीचर एक्सट्रेक्शनः

शब्द भेद (POS) टेगिंग : वाक्य में प्रयुक्त शब्दों को उनके शब्द भेदों से जोड़ने की प्रक्रिया को शब्द भेद टेगिंग कहा जाता है। इससे हमें वाक्य की व्याकरणिक संरचना का पता चलता है।

उदाहरण

वाक्यः मैं शराब को मना करता हूँ पर वो मेरी सुनती ही नहीं।

शब्द भेद (POS) टेगिंग : मैं _PRP शराब _NN को _ PSP मना _VM करता _VAUX हूँ _VAUX पर _PSP वो _PRP मेरी _NNP सुनती _VM ही _RP नहीं _NEG । _SYM

बैग ऑफ वड्र्स : प्रकृतिक भाषा की अव्यवस्थितता कंप्यूटर के लिए बेहद बड़ी समस्या होती है। मशीन लर्निंग तकनीक अव्यवस्थित टेक्स्ट के साथ प्रयोग करने में असमर्थ होती है। इस कारण हमें टेक्स्ट को वेक्टर रूप में बदलने की आवश्यकता होती है। 'बैग ऑफ वड्र्स' तकनीक हमें इसमें मदद करता है। हमने इसे पाइथन प्रोग्रामिंग द्वारा लागू किया है।

उदाहरण

दस्तावेज़

मैं आपके साथ सहमत होता लेकिन फिर हम दोनों हीं गलत होंगे।

मैं आपका अपमान नहीं कर रहा हूँ मैं आपका वर्णन कर रहा हूं।

इस दुनिया में सबसे असामान्य चीज़ सामान्य ज्ञान है।

मेरे द्वारा कही गई मतलबी भयानक सटीक बातों के लिए क्षमा करें।

मुझे लिखना पसंद है बस तकलीफ कागजों से है।

वेक्टर

[[0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1]
[1 0 1 1 0 2 0 0 0 0 0 1 0 0 1 0 0 0 0 2 1 0 0 0 0 0]
[0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]
[0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 1 1 0 0 0]
[0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0]]

शब्दावली

{'हम' : 25, 'आप': 2, 'सट': 22, 'वर': 20, 'इस': 4, 'सहमत': 24, 'कह' : 6, 'सबस': 23, 'मतलब': 18, 'षम': 21, 'अस': 1, 'रह': 19, 'णन': 11, 'गलत': 10, 'पस': 15, 'कर': 5, 'नक': 13, 'आपक': 3, 'अपम': 0, 'नह': 14, 'गय': 9, 'भय': 17, 'तकल': 12, 'खन': 7, 'बस': 16, 'गज': 8}

परिणाम एवं विवेचना

यह प्रयोग कार्य हिन्दी के डेटासेट पर किया गया है। डेटासेट को 80:20 अनुपात में बांटा गया है जिसमें से 80% डेटा को ट्रेनिंग के लिए तथा 20% डेटा को टेस्टिंग के लिए इस्तेमाल किया गया है। इसे पाइथन प्रोग्रामिंग की एस. के. लर्न. लाइब्रेरी की सहायता से अनियमित तरीके से बांटा गया है (सारणी 2)।

हमारे वर्गीकरण प्रयोग के लिए हमने मशीन लर्निंग की चार प्रचलित वर्गीकरण तकनीकों का इस्तेमाल किया हैः नेव बेज़, सपोर्ट वेक्टर मशीन, डिसीजन ट्री तथा न्यूरल नेटवर्क। इन चारों के कनफ्यूशन मैट्रिक्स कुछ इस प्रकार हैं (सारणी 3)।

इन परिणामों के मूल्यांकन के लिए हमने इनकी प्रिसिशन एरिकॉल तथा एफ़-मेज़र के मान निकाले, जो कुछ इस प्रकार हैं (सारणी 4)।

प्रयोग परिणामों के अनुसार, डिसीजन ट्री का वर्गीकरण 72% प्रिसीशन के साथ बाकी तीन वर्गीकरण तकनीकों के मुक़ाबले सबसे बेहतर परिणाम देता है जबकि न्यूरल नेटवर्क 68% प्रसीजिन के साथ दूसरे स्थान पर रहा।

सारणी 2 शब्दकोश में वाक्यों को विभाजित करने की नीति

प्रकार	ट्रेनिंग	टेस्टिंग
तंज़	400	100
सामान्य	400	100

सारणी 3 (क) नेव बेज़, (ख) सपोर्ट वेक्टर मशीन, (ग) डिसीजन ट्री तथा (घ) न्यूरल नेटवर्क की कन्युशन मैट्रिक्स

	अनुमान			अनुमान			अनुमान			अनुमान	
असल	58	52	असल	68	64	असल	62	77	असल	60	73
	48	42		36	32		23	38		27	40
	क.			ख.			ग.			घ.	

सारणी 4 नेव बेज़, सपोर्ट वेक्टर मशीन, डिसीजन ट्री तथा न्यूरल नेटवर्क के प्रिसिशन एरिकॉल तथा एफ़-मेज़र के मान

	प्रिसिशन	रिकॉल	एफ़-मेज़र
नेव बेज़	0.54	0.52	0.52
सपोर्ट वेक्टर मशीन	0.65	0.51	0.57
डिसीजन ट्री	0.72	0.44	0.54
न्यूरल नेटवर्क	0.68	0.45	0.54

निष्कर्ष

तंज़ वाक्यों को उनके अस्पष्ट व्यवहार के लिए जाना जाता है। इस कारण इन्हें पहचान पाना एक चुनौती से कम नहीं। आम तौर पर काम में लिए जाने वाली भावना विश्लेषण तकनीकें इसे पहचान पाने में असमर्थ होती हैं और इसी वजह से इनका गलत वर्गीकरण कर देती हैं। इस कारण से परिणाम सही नहीं मिलते। ऐसे वाक्यों से उसके छुपे हुए अर्थ को खोज कर निकालना मुश्किल था। इस प्रयोग कार्य ने एक कोशिश की है तंज़ को सामान्य वाक्यों के मुक़ाबले पहचानने की और सफलतापूर्ण निष्कर्ष भी दिये हैं। इस प्रयोग ने मशीन लर्निंग के वर्गीकारकों को काम में लिया और डिसीजन ट्री का वर्गीकरण सबसे भरोसेमंद पाया। आगामी कार्यों के लिए सलाह के तौर पर- शब्दकोश को और बढ़ाने से परिणामों में महत्वपूर्ण सुधार हो सकता है; साथ ही, हिन्दी भाषा के लिए संसाधनों की कमी के कारण हम कई पहलुओं पर काम नहीं कर पाये, उनके विकास के लिए काम किया जा सकता है।

संदर्भ

1. Joshi A, Balamurali A R & Bhattacharyya P, A Fall-back strategy for Sentiment Analysis in Hindi: A Case Study. Proceedings of the 8th ICON (2010).
2. Vellore, V.I.T. Context-based Sarcasm Detection in Hindi Tweets. In Meeting on Human Language Technologies. ACL (2011) 581-586.
3. Desai N, & Dave A D, Sarcasm detection in Hindi Sentences using Support Vector Machine, *International Journal*, **4**(7), (2016) 8-15.
4. Paul S, Joshi N & Mathur I, Development of a hindi lemmatizer. arXiv preprint arXiv: (2013) 1305.6211.
5. Ramanathan A & Rao D D, A lightweight stemmer for Hindi. In the Proceedings of EACL (2003, April).