सीमित लंबाई के सोशल मीडिया के लिए स्वचालित स्पैमर डिटेक्शन

Automated Spammer Detection for Limited Length Social Media

डॉ शिल्पा मेहता
Dr. Shilpa Mehta
Professor ECE – SoE, Presidency University Bangalore.
shilpamehta1.official@gmail.com, shilpamehta@presidencyuniversity.in

सारांशः

आज विभिन्न प्रसिद्ध सूचना साझाकरण और सामाजिक नेटवर्किंग सेवाएं जैसे ऑरकुट, ट्विटर, फेसबुक, आदि उपयोगकर्ताओं को संदेशों का आदान–प्रदान करने की अनुमति देती हैं। ट्विटर संदेशों को एक निश्चित लंबाई तक सीमित करता है। इसकी प्रकृति एक खुली हुई प्रकृति है, और इसके उपयोगकर्ताओं का विश्व भर में एक विशाल उपयोगकर्ता आधार है। स्पैमर इस सुविधा का उपयोग करते हैं। अफवाह फैलाना, साइबर हमला करना, ट्रोल करना और पीछा करना जैसे साइबर अपराध आम हैं। यहां प्रस्तावित तरीके उपयोगकर्ताओं को अपने अनुयायियों के साथ बातचीत के आधार पर चिह्नित करने की कोशिश करते हैं। स्पैमर स्वयं अपनी गतिविधियों को समायोजित करके पारंपरिक तरीकों को मूर्ख बना सकते हैं। लेकिन यह विधि उपयोगी साबित होती है, क्योंकि उपयोगकर्ता अपनी गतिविधियों को नियंत्रित कर सकता है, लेकिन अनुयायियों की गतिविधियों और सुविधाओं को नियंत्रित करना उनके लिए संभव नहीं है। तीन क्लासिफायर, डिसिशन ट्री, बायेसियन नेटवर्क, और रैंडम फ़ॉरेस्ट का उपयोग उपयोगकर्ता के बारे में सीखने और ट्वीट के लक्षण पहचानने के लिए किया गया। अध्ययन दर्शाता है कि इंटरएक्टिव विशेषताएं और समुदाय–आधारित विशेषताएं स्पैम उपयोगकर्ताओं की पहचान करने में अच्छी साबित होती हैं, मेटाडेटा आधारित निर्णयों की तुलना में उनका प्रदर्शन बेहतर पाया गया है।

Abstract:

Today, various famous social networking services like Orkut, Twitter, Facebook, and many others, openly and freely allow users to exchange messages among themselves. In this paper, we talk of Twitter, which restricts the messages to a character limit. It has an open nature and a huge user base, and spammers utilize this characteristic. Cybercrimes are quite commonplace. A few commonly known ones are, spreading rumors, cyberbullying, trolling, and stalking. The approaches proposed in this work try to characterize users for the detection of such unwanted things. While traditional approaches try to do this based on the content of the messages, the approaches discussed here work based on the interactions of users with their followers along with the content. This is extremely useful, as a user who is spamming can easily control one's own activities, but controlling the activities of ones' network (followers) is not possible. Three classifiers are used for learning and characteristic identification. They are Decision Tree, Bayesian network, and Random Forest. Dataset from both genuine real users and spammers was used for testing. Study indicates that Interactive characteristics and community-based characteristics prove good in identifying spam users. They perform better than Metadata based decisions.

संकेत शब्दः ट्विटर, स्पैम, सोशल मीडिया, साइबर अपराध

Keywords: Twitter, Spam Detection, Social Media, Cybercrime

1. विषय परिचयः Introduction

ट्विटर और अन्य सोशल मीडिया प्लेटफॉर्म, अपने उपयोगकर्ताओं को दुनिया के लिए वास्तविक समय (Real Time) में विभिन्न विषयों और रुचियों की जानकारी उपलब्ध कराते हैं। ऐसे प्लेटफ़ॉर्म ऑनलाइन सोशल नेटवर्क (OSN) कहलाते हैं, जो उपयोगकर्ताओं को अपनी–अपनी रुचि की चीज़ों को साझा करने में सक्षम बनाते हैं। इसमें समाचार, विभिन्न चीजों के बारे में उनकी अपनी राय, पारिवारिक आउटिंग और अन्य तस्वीरें और वीडियो, समारोह और उनके मूड शामिल हैं। वे अपने जानकार अन्य उपयोगकर्ताओं के साथ विभिन्न विषयों पर तर्क और चर्चा करते हैं, जैसे राजनीति, महत्वपूर्ण घटनाएँ, वर्तमान विश्व परिस्थितियाँ और इसी तरह अन्य विषय भी। जब कोई उपयोगकर्ता कुछ भी ट्वीट करता है, तो अनुयायी उसे आमतौर पर पढ़ते हैं और पुनः ट्वीट भी करते हैं। इससे जानकारी का व्यापक फैलाव हो पाता है। OSNs ऑनलाइन प्लेटफॉर्म पर उपयोगकर्ताओं के व्यवहार के अध्ययन और विश्लेषण की आवश्यकता होती है।

ट्विटर की शुरुआत 2006 में हुई थी। इसमें अक्षर सीमा की नीति है। आम तौर पर एक ट्वीट में अधिकतम 280 वर्ण हो सकते हैं। उपयोगकर्ता अपनी पसंद के अनुसार अन्य लोगों (प्रसिद्ध हस्तियों सहित) का अनुसरण कर सकते हैं। समाचार चैनलों का अनुसरण करने और सदस्यता लेने के विकल्प भी उपलब्ध हैं। प्रत्येक पंजीकृत अनुयायी को सदस्यता खातों की अपडेट प्राप्त होती है। ट्विटर और फेसबुक जैसे OSN मूल रूप से सामाजिक उद्देश्यों के लिए बनाए गए थे, लेकिन वे खुले हैं और उनके विशाल जनाधार हैं। उनके पास संदेश साझा करने के लिए आरामदायक और आसान विकल्प हैं। ये दो विशेषताएं "सामाजिक बॉट" और साइबर अपराधियों को आकर्षित करती हैं।

OSNs नए परिष्कृत और जटिल हमलों और खतरों, जैसे साइबर बदमाशी, गलत सूचना फैलाना, कट्टरपंथीकरण और विभिन्न अन्य अवैध गतिविधियों में प्रवीण हैं। फ़िशिंग (phishing) और स्पैमिंग जैसे आम पारंपरिक साइबर हमलों की भी समस्या है। जैसे–जैसे इन्हे पकड़ने के तरीके विकसित हुए, वैसे–वैसे उनके साथ ही समानांतर रूप से पता लगाने से बचने के लिए होशियार संस्करण विकसित हुए। रिपोर्टों का कहना है कि उपयोगकर्ताओं के साथ–साथ ट्वीट का एक बड़ा प्रतिशत क्रमशः बॉट और स्पैम है। स्पैम बॉट्स पहले तो दुसरे उपयोगकर्ताओं के विश्वासपात्र बनने के लिए मानवीय व्यवहार की नकल करते हैं, और फिर इसका उपयोग अनचाही गतिविधियों के लिए करते हैं। (1)

2. साहित्य सर्वेक्षण : Literature Review

उद्योग जगत के एक्सपर्ट और शिक्षाविद, दोनों प्रकार के शोधकर्ता साइबर अपराधियों को हराने के लिए काम कर रहे हैं, और उपयोगकर्ताओं को OSNs पर एक सुरक्षित और सुखद अनुभव देने का प्रयास कर रहे हैं। विभिन्न स्पैम डिटेक्शन एप्रोच सभी प्लेटफार्मों पर उपलब्ध हैं, और नए एप्रोच लगातार प्रस्तावित हो रहे हैं (2, 3)। दुर्भाग्य से, पता लगाने की तकनीक के साथ साथ ही पता लगाने से बचने के लिए तकनीक विकसित होती है। अधिकांश तरीके उपयोगकर्ता के स्वयं के प्रोफाइल और गतिविधियों के लक्षणों का उपयोग करते हैं। इसलिए स्पैमर ऐसी कमजोरियों का उपयोग करके पकड़े जाने से बच जाते हैं। अधिकांश उपयोगकर्ता जो एक–दूसरे से बातचीत करते हैं, वे एक–दूसरे की वास्तविक पहचान जानते हैं। वास्तविक लोग असल जीवन में अपने पड़ोस, मित्र मंडली, कार्यालय मंडल और व्यक्तिगत पसंद–नापसंद के अनुसार विभिन्न समुदायों से संबंध रखते हैं, और एक दूसरे का अनुसरण करते हैं। इसके विपरीत, स्पैमर अनियमित उपयोगकर्ताओं का अनुसरण करते हैं, जिसके कारण उन्हें बहुत कम पारस्परिकता मिलती है। इसके अतिरिक्त, ऐसे अनुयायियों के बीच कम पारस्परिक कनेक्शन होते हैं, जो बातचीत को और साथ ही समुदाय

आधारित विशेषताओं को कम करते हैं। वास्तविक उपयोगकर्ताओं और स्पैमर्स के बीच यह अंतर हमें स्पैमर्स का पता लगाने में उपयोग की जाने वाली एक विधि प्रदान करता है। इस तरह के पहचान तंत्र से बचने के लिए, स्पैमर एक–दूसरे का अनुसरण करते हैं, और नकली समुदाय बनाते हैं। लेकिन यह वास्तविक लोगों को स्पैम करने के उनके मूल उद्देश्य को हरा देगा, क्योंकि वे स्पैमिंग समूहों के भीतर रहने के लिए प्रतिबंधित हो जाते हैं। इसके अलावा, इन नकली समुदायों के सदस्यों का बड़ा प्रतिशत स्पैमर्स का होगा और समान व्यवहार होंगे, जिससे समग्र का पता लगाने की संभावना बढ़ जाती है।

स्पैम ने उन दिनों से समस्याएँ पैदा की हैं जब इंटरनेट की शुरुआत हुई थी। इन्हें 1978 में ARPANET में रिपोर्ट किया गया था। लेकिन उस समय, किसी को इसकी चिंता नहीं थी। ई–मेल बस शुरुआत कर रहे थे, और स्पैम को अभी तक मान्यता नहीं मिली थी। लेकिन यह समस्या समय के साथ और तेज होती गई, और यह आज एक विशाल संकट बन गई है। अनचाहे स्पैमिंग (स्पैम ईमेल / सोशल बॉट्स / और स्पैम्बोट्स) के विभिन्न रूपों को पकड़ने में सक्षम कई तकनीकों को बनाया गया है। शोध पत्र (4) में, पहचान की धोखे की रोकथाम तकनीक प्रस्तावित है। यह सामान्य योगदान नेटवर्क डेटा का उपयोग करता है और उप–समुदाय में प्रवेश करने की कोशिश कर रहे भ्रामक खातों की पहचान करता है। धोखे का तात्पर्य है, किसी ऐसे व्यक्ति को झूठी जानकारी जानबूझकर हस्तांतरित करना, जो झूठ से अनजान है। यह एक पहचान और रोकथाम तंत्र दोनों के रूप में प्रभावी प्रतीत होता है। लेकिन इसकी दक्षता काफी कम है, खासकर बड़े नेटवर्क में तो बहुत ही कम है।

WARNINGBIRD (5) को स्पैमर्स के ट्विटर स्ट्रीम में पाए जाने वाले संदिग्ध URL को वर्गीकृत करने और "लगभग" वास्तविक समय ऑपरेटिंग सिस्टम के रूप में कार्य करने के लिए डिज़ाइन किया गया था। इसका उद्देश्य केवल ट्वीट्स को वर्गीकृत करना है, न कि उन पृष्ठों की जांच करना,जो वहां से खुल रहे हैं। सहसंबंध (correlation) को कई ट्वीट्स से निकाला जा सकता है। हमलावर के पास भी तो असीमित संसाधन नहीं हैं और इसलिए URL का अनिवार्य रूप से पुनः उपयोग किया जाता है। ऐसी सहसंबद्ध या साझा URL पुनर्निर्देशित श्रृंखलाओं का पता लगाने में अच्छी तरह से काम करता है। इस तकनीक की मुख्य सीमाएँ इसकी धीमी गति और संसाधनों का अप्रभावी और कम उपयोग हैं। शोध पत्र (6) स्पैम जांच के लिए एक विधि प्रस्तावित करता है, जो ट्विटर की अपनी स्पैम नीति का उपयोग करता है। बायेसियन वर्गीकरण एल्गोरिथ्म (6) सामान्य और संदिग्ध प्रकार के व्यवहारों को अलग करता है। ग्राफ़–आधारित सुविधाएँ भी उपयोगी पायी जाती हैं, क्योंकि स्पैम खाते आवश्यक रूप से बड़ी संख्या में उपयोगकर्ताओं का अनुसरण करते हैं। जब हम सामग्री आधारित विशेषताओं की जांच करते हैं, तो स्पैम खातों में आमतौर पर कई डुप्लिकेट ट्वीट पाए जाते हैं, जो इसकी पहचान का कारण बन जाता है। हालांकि, वास्तविक उपयोगकर्ता बार–बार ट्वीट पोस्ट कर सकते हैं, इसलिए यह एक प्रामाणिक विधि नहीं है।

3. प्रस्तावित प्रणालीः एक सम्मिश्रित दृष्टिकोण The Proposed Technique: A Hybrid Approach

प्रस्तावित कार्यप्रणाली विभिन्न तकनीकों के संयोजन पर काम करती है, जिसमें सामग्री, सामुदायिक विशेषताएं और इंटरैक्शन शामिल हैं। नेटवर्क श्रेणी को समुदाय–आधारित और इंटरैक्शन प्रकारों के लिए उप–वर्गीकृत किया गया है, जो इंटरैक्शन नेटवर्क से आते हैं। मेटाडेटा विशेषताएँ उपयोगकर्ताओं के स्वयं के ट्वीट से आती हैं, और सामग्री–आधारित विशेषताएँ पोस्ट की गुणवत्ता और पोस्टिंग व्यवहार का उपयोग करती हैं। स्पैमर सामग्री और मेटाडेटा आधारित पहचान से बचने के लिए स्वयं अपने पोस्ट तो समायोजित कर सकते हैं, लेकिन अनुयायियों और समुदाय–आधारित विशेषताओं को बायपास करना मुश्किल होता है।

3.1 विधि और डाटासेट : Methodology and Dataset

हम प्रयोगात्मक विश्लेषण और मूल्यांकन के लिए, वास्तविक उपयोगकर्ताओं और स्पैमर्स दोनों के ट्विटर डेटासेट के उपयोग पर चर्चा करते हैं। प्रत्येक उपयोगकर्ता की प्रोफ़ाइल जानकारी के साथ साथ ही उसके अनुयायी और उसकी अनुसरण सूची, और ट्वीट का विवरण भी उपलब्ध होता है। हम अनुयायियों वाले उपयोगकर्ताओं का डाटासेट लेकर काम करेंगे। प्रतीकों के लिए तालिका 1 देखें:

Symbol	Description
$\overleftarrow{u}$	Follower set of user u (set of users that follow u)
$\overrightarrow{u}$	Following set of user u (set of users that are followed by u)
$N(u)$	Total number of tweets tweeted by user u
$u_{\overleftarrow{v}}$	A follower, named v, of user u
$\overrightarrow{u_{\overleftarrow{v}}}$	Following set of the follower v of user u

तालिका 1: प्रतीक विवरण

हम अपनी तकनीकों को निम्न लिखित मुख्य गुणों पर आधारित करेंगे:

3.2 मेटाडेटा–आधारित व्यवस्था Metadata Based Techniques

3.3 सामग्री–आधारित विशेषताएँ Content-Based Characteristics

3.4 पारस्परिक विचार–विमर्श आधारित गुण: Interaction-Based Characteristics

3.5 समुदाय आधारित विशेषताएँ Community-Based Characteristics

3.2 मेटाडेटा–आधारित व्यवस्था Metadata Based Techniques:

मेटाडेटा ट्वीट विशेषताओं का वर्णन करने वाली जानकारी का प्रतिनिधित्व करता है, और सूचना स्रोत खोजने में उपयोगी हो सकता है। इन चार विशेषताओं को पहचान के लिए उपयोग किया जाता है।

3.2.1 रीट्वीट का अनुपात Retweet Ratio (RR)

स्वचालित स्पैमर मनुष्यों की तरह ट्वीट नहीं करते। बॉट सामान्य रूप से दूसरों के ट्वीट को रीट्वीट करते हैं, या डेटाबेस से ट्वीट का उपयोग करते हैं, या फिर संभाव्य तरीकों का उपयोग करके ट्वीट बनाते हैं। इस का मूल्यांकन RR के साथ किया जा सकता है, जो कि के उपयोगकर्ता द्वारा दिए गए रीट्वीट का कुल ट्वीट पर अनुपात है। यह वास्तविक लोगों के लिए कम होना चाहिए जबकि यह स्पैमर के मामले में उच्च होगा। गणितीय रूप से,

RR(u) = {RT(u)} / {N(u)}, *(Eq 3.2.1)*

जहाँ RT(u) उस उपभोक्ता के कुल रीट्वीट और N(u) उस उपभोक्ता के कुल ट्वीट है।

3.2.2 स्वचालित ट्वीट अनुपात Automated Tweet Ratio (AR)

कुछ स्पैमिंग खाते OSNs द्वारा प्रदान किए गए API का उपयोग करते हैं। ट्विटर में एक सार्वजनिक एपीआई है, और इसका उपयोग कई स्पैमर्स द्वारा किया जाता है। अपंजीकृत और तृतीय–पक्ष एप्लिकेशन से उत्पन्न ट्वीट्स को स्वचालित ट्वीट कहा जाता है। किसी दिए गए उपयोगकर्ता का AR, उस के एपीआई से ट्वीट और कुल ट्वीट्स का अनुपात है।

AR(u) = {A(u)} / {N(u)}, *(Eq 3.2.2)*

जहां A(u) एपीआई का उपयोग करते हुए ट्वीट्स की संख्या है। वास्तविक लोगों के लिए AR कम होना चाहिए, जबकि स्पैमर के लिए अधिक।

3.2.3 ट्वीट टाइम स्टैंडर्ड डेविएशन Tweet Time Standard Deviation (TSD)

स्वचालित स्पैमर अपनी गतिविधि का समय निर्धारित करने के लिए रैंडम नंबर जनरेटर का उपयोग करते हैं। बॉट टाइम एक्टिवेशन फ़ंक्शन के अनुसार समय में एक निर्दिष्ट बिंदु पर सक्रिय हो जाते हैं। हालाँकि वे एल्गोरिदम भी कुछ नियमों का पालन करते हैं। ज्क्ष्क ट्वीट समय भिन्नताओं को कैप्चर करता है। गणितीय रूप से:

$$TSD(u) = \frac{\sum_{i=1}^{N(u)} (t_i - \bar{t})^2}{N(u)} \quad (Eq\ 3.2.3)$$

जहां ti iवीं ट्वीट समय है, ट्वीट्स के बीच औसत समय $\bar{t}$ है, और N (u) उपभोक्ता u द्वारा किए गए कुल ट्वीट हैं। स्वचालित स्पैमर में कम टीएसडी होगा जबकि वास्तविक लोगों में पास अधिक।

3.2.4 ट्वीट समय अंतराल मानक विचलन Tweet Time Interval Standard Deviation (TISD)

TISD लगातार गतिविधियों में पैटर्न का उपयोग करता है। बॉट्स द्वारा किए गए ट्वीट में नियमित अंतराल होता है जबकि मनुष्यों में अनियमित अंतराल होता है। गणितीय रूप से,

$$TISD(u) = \frac{\sum_{i=1}^{n} (T_i - \bar{T})^2}{N(u)} \quad (Eq\ 3.2.4)$$

जहां T1, T2, ..., Tn लगातार ट्वीट्स के बीच समय अंतराल देते हैं और औसत समय अंतराल है। स्वचालित स्पैमर्स के लिए यह भी कम होगा।

3.3 सामग्री–आधारित विशेषताएँ Content-Based Characteristics

वर्तमान विधियां सामग्री गुणवत्ता का उपयोग एक संकेतक के रूप में भी करती हैं। स्पैमर्स उपयोगकर्ताओं को लुभाने और धोखा देने के लिए ट्वीट पोस्ट करते हैं, और विशिष्ट विशेषताओं का भी उपयोग किया जा सकता है। जैसे किः

3.3.1 URL अनुपात URL Ratio (UR)

उपयोगकर्ता आमतौर पर अपनी रुचि के विचार और विचार पोस्ट करते हैं, और समाचार लेख और कहानियां साझा करते हैं। इस तरह के ट्वीट में कभी –कभी तो संबंधित पृष्ठ के स्रोत URL हो सकते हैं, लेकिन हमेशा नहीं होंगे। RR का समीकरण देखिए

UR(u) = {U(u)} / { N(u)}, *(Eq 3.3.1)*

जहां U(u) ट्वीट में उपयोग किए गए URL के साथ ट्वीट्स की संख्या है, और N(u) यू द्व ारा किए गए कुल ट्वीट हैं। स्पैमर्स को URL का उपयोग करना आवश्यक है, ताकि वे जो वो चाहते हैं, वह करने में सक्षम हो सकें, ताकि वे सामान्य उपयोगकर्ताओं को अपने लक्षित पृष्ठों पर खींच सकें। इसके लिए, Spammers को URL का उपयोग करना ही होता है। इसलिए उनके ज्यादातर ट्वीट्स में URL होते हैं, और इसलिए UR मान लगभग 1 हो जाता है। इसके विपरीत, वास्तविक उपयोगकर्ताओं के UR छोटे (0 के करीब) होते हैं, क्योंकि अधिकांश समय वे अपनी भावनाओ को व्यक्त कर रहे होते हैं और यूआरएल पोस्ट नहीं करते। यह भी वास्तविक उपयोगकर्ता से स्पैमर का पता लगाने का एक और तरीका है।

3.3.2 अद्वितीय URL अनुपात न्दपुनम URL Ratio (UUR)

URL का अत्यधिक उपयोग अपने आप में संदेहास्पद है, लेकिन एक ही URL के बार–बार दोहराया जाने से संदेह की तीव्रता बढ़ जाती है। स्पैमर्स बार–बार एक ही URL का उपयोग करके उपयोगकर्ताओं को फंसाने की कोशिश करते हैं, उन्हें दुर्भावनापूर्ण साइट पर पुनर्निर्देशित करने के लिए। उपयोगकर्ताओं के ट्वीट में URL की विशिष्टता के लिए अद्वितीय URL अनुपात का उपयोग करके इसकी पहचान की जाती है।

UUR(u) = {UU(u)} / {U(u)}, *(Eq 3.3.2)*

जहां UU (u) अद्वितीय URL है और U(u) ट्वीट्स में कुल URL हैं। यह स्पैमर के लिए लगभग 1 होगा लेकिन वास्तविक उपयोगकर्ताओं के लिए कम मूल्य होगा, जिससे स्पैमर का पता लगाया जा सकेगा।

3.3.3 उल्लेख अनुपात Mention Ratio (MR)

ट्विटर हैंडल "@userid" का उपयोग करके उपयोगकर्ताओं को ट्वीट्स में दूसरों को टैग करने की अनुमति देता है। स्पैमर्स ट्वीट में उपयोगकर्ता नामों का उल्लेख करके इस सुविधा का उपयोग करते हैं, एवं उन्हें संदेश भेजने वाले के बारे में जानने के लिए क्लिक करने के लिए उकसाते हैं।

MR(u) = {M(u)} / N(u), *(Eq 3.3.3)*

जहाँ M(u) उल्लेखों की गिनती है। यह पैरामीटर भी वास्तविक लोगों के लिए कम और स्पैम बॉट के लिए उच्च होना अपेक्षित है।

3.3.4 अद्वितीय उल्लेख अनुपात Unique Mention Ratio (UMR)

वास्तविक उपयोगकर्ताओं के वास्तविक लोगों के साथ कई संबंध हो सकते हैं लेकिन बातचीत सभी के साथ नहीं होती है। यही कारण है कि कोई भी वास्तविक उपयोगकर्ता अपने सभी परिचितों को टैग नहीं करता है। इसके विपरीत, स्पैमर किसी को भी अनियमितरूप से टैग करते हैं। गणितीय रूप से, यूएमआर अनुपात निम्नलिखित समीकरण द्वारा निर्धारित किया जाता है:

UMR(u) = {UM(u)} / {M(u)}, *(Eq 3.3.3)*

जहां UM(u) उपयोगकर्ता यू द्वारा किये गए अद्वितीय उल्लेख है, और M(u) यू द्वारा किये गए कुल उल्लेख हैं। वास्तविक लोगों के लिए UMR कम है क्योंकि वे विशिष्ट लोगों के साथ बातचीत करते हैं। यह स्पैम के लिए उच्च होना अपेक्षित है।

3.4 पारस्परिक विचार–विमर्श आधारित गुणः Interaction-Based Characteristics:

OSNs से उपलब्ध उपयोगकर्ता इंटरैक्शन डेटा हमें उचित निर्णय लेने के लिए कई स्तरों पर बहुत सारी जानकारी प्रदान करता है। यह धोखाधड़ी का पता लगाने, उपयोगकर्ता की वास्तविक दुनिया की पहचान, ग्राहक व्यवहार विश्लेषण और उपयोगकर्ता के व्यवहार का पूर्वानुमान लगाने में हमारी मदद करता है। कोई भी उपयोगकर्ता दूसरों की सदस्यता / गतिविधियों का पालन अपनी ओर से कर सकता है, लेकिन दूसरों को उसका अनुसरण करने के लिए मजबूर नहीं कर सकता है। यह हमें और अधिक पहचान कर पाने की तरकीबें / तकनीक प्रदान करता है। इसके कई प्रकार हैं जो नीचे समझाए गए हैं।

3.4.1 अनुयायी अनुपात Follower Ratio (FR)

एक उपयोगकर्ता के अनुयायियों की गिनती अन्य उपयोगकर्ताओं के बीच उसके विश्वास स्तर को इंगित करती है। ओएसएन नेटवर्क में जुड़े लोग आम तौर पर वास्तविक दुनिया में भी एक दूसरे को जानते हैं (सेलिब्रिटी / लोकप्रिय उपयोगकर्ताओं को छोड़कर)। इसलिए, वास्तविक लोगों के लिए फॉलो–बैक दर सामान्य रूप से अधिक होती है। उपयोगकर्ता u का FR नेटवर्क में उस का अनुयायी अनुपात है। FR (फॉलो–बैक दर) वास्तविक उपयोगकर्ताओं के लिए उच्च और स्पैमर्स के लिए कम होती है।

$$FR(u) = \frac{|\overleftarrow{u}|}{|\overrightarrow{u} \bigcup \overleftarrow{u}|} \qquad (Eq\ 3.4.1)$$

3.4.2 प्रतिष्ठा और पारस्परिकता Reputation and Reciprocity

समाज में लोगों की वास्तविक विश्व में प्रतिष्ठा उनके समुदाय में उनके विचारों और उनके प्रति लोगों के विश्वासों को इंगित करती है, और आभासी दुनिया में भी यही लागू होता है। इसका मतलब है कि यदि A ट्विटर पर B का अनुसरण करता है, तो यही अपेक्षित होगा कि B भी A का अनुसरण करे, जिससे A के लिए उच्च पारस्परिकता दर हो। पारस्परिकता दर नेमत के किसी का अनुसरण करने के बाद, वापस अनुसरण के अनुपात को इंगित करता है।

$$R(u) = \frac{|\overleftarrow{u} \cap \overrightarrow{u}|}{|\overrightarrow{u}|} \qquad \text{(Eq 3.4.2)}$$

R(u) आम तौर पर स्पैमर्स के लिए कम और वास्तविक उपयोगकर्ताओं के लिए उच्च होना अपेक्षित है।

3.4.3 अनुयायी–आधारित प्रतिष्ठा Follower-Based Reputation (FBR)

एक उपयोगकर्ता की प्रतिष्ठा उसके साथ जुड़े उपयोगकर्ताओं से विरासत में मिली हो सकती है। इस विशेषता का उपयोग, उपयोगकर्ता के अनुयायियों की प्रतिष्ठा से उसकी प्रतिष्ठा तय करने के लिए किया जाता है। FBR दिए गए उपयोगकर्ताओं के अनुयायियों की औसत प्रतिष्ठा है। गणितीय रूप से,

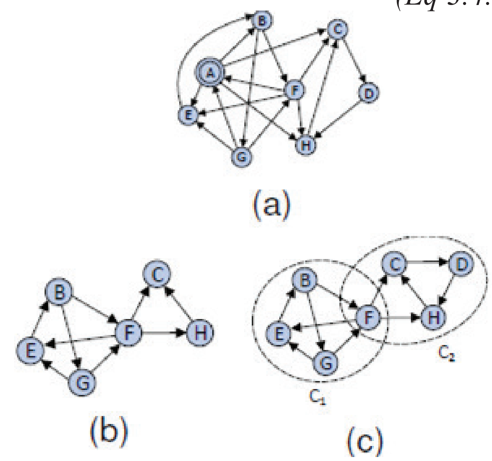
$$FBR(u) = \frac{\sum_{u_{\overleftarrow{v}} \in \overleftarrow{u}} R(u_{\overleftarrow{v}})}{|\overleftarrow{u}|} \qquad \text{(Eq 3.4.3)}$$

जहाँ R(u) अनुयायियों की प्रतिष्ठा है, FBR आम तौर पर स्पैमर्स के लिए कम और वास्तविक उपयोगकर्ताओं के लिए उच्च होना अपेक्षित है।

3.4.4 क्लस्टरिंग गुणांक Clustering Coefficient (CC)

नेटवर्क नोड का CC, उस विशेष नोड से गुजरने वाले कनेक्शनों को छोड़कर, आसन्न नोड्स के इंटरकनेक्टिविटी के घनत्व का प्रतिनिधित्व करता है। यह स्वयं पड़ोसियों के बीच विश्वास स्तर को इंगित करता है भले ही यह उपयोगकर्ता मौजूद न हो। क्योंकि वास्तविक लोगों के वास्तविक नेटवर्क हैं और सभी एक दूसरे से परस्पर परिचित हैं, वास्तविक दुनिया में जुड़े उपयोगकर्ताओं के बीच विश्वास होने पर वास्तविक उपयोगकर्ता के पास उच्च CC (1 के करीब) हो सकता है। एक स्पैमर नेटवर्क की तुलना में उनका नेटवर्क घना है।

$$CC(u) = \frac{E_u}{K_u \times (K_u - 1)} \qquad \text{(Eq 3.4.4)}$$



चित्र 3.4.4 (अ) "ए" का नेटवर्क (ब) अ के न होते हुए उसके सम्बंधित बिंदुओं का आपसी नेटवर्क, (स) "ए" के पड़ोसियों के बीच एक समुदाय का नमूना।

Fig. 3.5 **समुदाय आधारित विशेषताएँ** Community - Based Characteristics

मानव प्राचीन काल से समाज में एक साथ रह रहे हैं। वास्तविक विश्व समुदायों में, उपयोगकर्ता एक दूसरे को जानते हैं और उनके बीच एक विश्वास स्तर है, और बाहरी लोगों की तुलना में आपस में उच्च संबंध घनत्व भी। अब हम इसके आधार पर कुछ दृष्टिकोणों को देखते हैं।

3.5.1 समुदाय–आधारित प्रतिष्ठा Community-Based Reputation (CBR)

किसी भी उपयोगकर्ता की प्रतिष्ठा उनके समुदायों की प्रतिष्ठा और उसके संबंधित सदस्यों की प्रतिष्ठा के अनुपात में है। वास्तविक उपयोगकर्ताओं के साथ जुड़े समुदायों में अच्छी प्रतिष्ठा हो, तो उसकी अपनी प्रतिष्ठा भी बढ़ जाती है। इस प्रकार, अच्छी प्रतिष्ठा वाले समुदायों में सम्मिलित होने भर से, उपयोगकर्ता की प्रतिष्ठा बढ़ जाती है यदि उपयोगकर्ता C1, C2, C3 के रूप में नामित समुदायों में मौजूद

है, तब उसका CBR मान यह होगाः

$$CBR(u) = \frac{\sum_{i=1}^{k}\left(\left(\sum_{j=1}^{|C_i|} R(C_i(j))\right) \big/ |C_i|\right)}{k} \quad (Eq 3.5.1)$$

जहां R (Ci (j)) ith समुदाय के रजी उपयोगकर्ता के लिए प्रतिष्ठा है। सीबीआर स्पैमर्स के लिए कम और वास्तविक उपयोगकर्ताओं के लिए उच्च है।

3.5.2 समुदाय आधारित क्लस्टरिंग गुणांक Community-Based Clustering Coefficient (CBCC)

यह गुणांक समुदायों के समूहों को इंगित करता है जिसमें उपयोगकर्ता मौजूद है। यदि CCi ith समुदाय का क्लस्टर गुणांक है, और उपयोगकर्ता U ऐसे K समुदायों का सदस्य है, तो U के लिए CBCC यह होगा,

$$CBCC(u) = \frac{\sum_{i=1}^{k} CC_i}{k} \quad , (Eq\ 3.5.2)$$

सीबीसीसी वास्तविक लोगों के लिए उच्च और स्पैमर्स के लिए कम होगा, क्योंकि स्पैमर्स से जुड़े स्वतंत्र उपयोगकर्ता शायद ही कभी अंतर–समुदाय होते हैं।

4. परिणाम और तुलना के लिए पैरामीटर Results and Parameters of Comparison

डेटासेट पर इस एल्गोरिथ्म के प्रदर्शन का विश्लेषण (और तुलना), मशीन लर्निंग के इन तीन तरीकों के साथ किया गयाः डिसिशन ट्री, रैंडम फारेस्ट और बैसिअन नेटवर्क। तुलना के लिए प्रयुक्त किये गए मीट्रिक थे अवास्तविक सकारात्मक या झूठी सकारात्मक दर false positive rate (FPR), खोज या पता लगाने की दर कमजमबजपवद तंजम (DR), और एफ–स्कोर F-Score।

$$DR = \frac{TP}{TP + FN}$$

$$DR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$F\text{-}Score = \frac{2 \times precision \times recall}{precision + recall}$$

TP (True Positive) टीपी वास्तविक सकारात्मकता संख्या को इंगित करता है, जो यह बताती है कि कितने असली स्पैमर्स को सही तरह से स्पैमर के रूप में वर्गीकृत किया गया। इसके विपरीत, एफ़ एन FN (False Negative) गलत नकारात्मकता है, जो यह बताती है कि कितने असली स्पैमर्स को गलती से असली उपयोगकर्ता समझ लिया। ठीक उसी तरह, एफपी FP (False Positive) झूठी सकारात्मकता संख्या है; यह वह संख्या है जो बताती है कि कितने वास्तविक लोगों को गलत तरह से स्पैमर के रूप में वर्गीकृत किया गया। और अंत में टीएन TN (true negatives) सच्चे नकारात्मकता संख्या हमें सही ढंग से वर्गीकृत वास्तविक उपयोगकर्ताओं की संख्या बताता है।

एफपीआर झूठी सकारात्मक दर है। FPR (false positive rate) एफपी (FP) और टीएन (TN) के संयोजन से प्राप्त होता है, जो कि कुल वास्तविक उपयोगकर्ताओं के अनुपात में स्पैमर्स के रूप गलत तरीके से वर्गीकृत वास्तविक उपयोगकर्ताओं का अंश है। एक अच्छे क्लासीफायर में कम थ्च्त्र होना चाहिए।

इसी तरह, डीआर (DR or recall rate or detection rate) टीपी (TP) और एफएन (FN) का संयोजन है। यह "पकडे गए " स्पैमर का कुल स्पैमर से अनुपात देता है। अच्छे क्लासिफायर में यह दर अधिक होनी चाहिए।

और अंत में, एफ–स्कोर (F Score) औसत सटीकता/ नीर क्षीर विवेकशीलता की शक्ति है, यह सभी पहचाने गए स्पैमर्स में से सही ढंग से पकडे गए (गलत पहचानों को छोड़ कर सिर्फ सही पहचाने गए स्पैमर्स की संख्या) स्पैमर्स का अनुपात है। यह

वर्गीकरणकर्ता की विभेदकारी शक्ति को दर्शाता है। एक उच्च एफ–स्कोर वाला क्लासिफायर वांछनीय है। ऊपर दिए गए समीकरण देखें। तालिका 2 इन मापदंडों के साथ विभिन्न एल्गोरिदम के प्रदर्शन की तुलना करके प्राप्त परिणामों को दिखाती है। अध्ययन यह भी बताते हैं कि, प्रदर्शन डेटासेट में "स्पैमर्स की गणना" के "वास्तविक उपयोगकर्ताओं की गणना" के अनुपात के साथ भी बदलता है। यह तालिका ३ में देखा जा सकता है।

Feature Set	Random Forest			Decision Tree			Bayesian Network		
	DR	FPR	F-score	DR	FPR	F-score	DR	FPR	F-score
F	0.97	0.01	0.97	0.94	0.04	0.94	0.90	0.01	0.94
F\Metadata Feature Set	0.96	0.03	0.97	0.93	0.05	0.94	0.92	0.02	0.94
F\Content Feature Set	0.95	0.02	0.96	0.92	0.05	0.93	0.90	0.04	0.94
F\Interaction Feature Set	0.93	0.02	0.95	0.93	0.05	0.93	0.85	0.04	0.89
F\Community Feature Set	0.94	0.02	0.95	0.93	0.05	0.93	0.84	0.02	0.90

तालिका 2: इन मापदंडों के साथ विभिन्न एल्गोरिदम के प्रदर्शन की तुलना के परिणाम

Spammers and Benign Users Ratio	random forest			Bayesian network			decision tree		
	DR	FPR	F-score	DR	FPR	F-score	DR	FPR	F-score
1:2	0.96	0.009	0.96	0.88	0.01	0.92	0.92	0.02	0.93
1:5	0.92	0.007	0.92	0.88	0.01	0.90	0.89	0.02	0.88
1:10	0.87	0.002	0.91	0.90	0.00	0.93	0.86	0.01	0.86

तालिका 3: वास्तविक उपयोगकर्ताओं और स्पैमर्स के विभिन्न अनुपातों के साथ एल्गोरिदम के प्रदर्शन में बदलाव

5. निष्कर्ष CONCLUSION

पुराने तरीके अपने प्रदर्शन और प्रोफाइल के आधार पर स्पैमर उपयोगकर्ताओं की पहचान करते हैं, जबकि प्रस्तावित दृष्टिकोण यह तो करता ही है, लेकिन इसके साथ ही इसके पड़ोसी नोड्स के आधार पर और इंटरैक्शन नेटवर्क के आधार पर भी काम करता है। मेटाडेटा–आधारित तकनीकों का प्रदर्शन उतना अच्छा नहीं होता है, क्योंकि वे विभिन्न एल्गोरिदम द्वारा मूर्ख बनाये जा सकते हैं। बातचीत–और समुदाय आधारित विशेषताओं को प्रयुक्त करने वाले तरीके और तरकीबें, ऊपर परिभाषित सभी मापदंडों के आधार पर बेहतर प्रदर्शन करती पायी गयीं हैं। आज जब सारी दुनिया अंतरजाल का प्रयोग करती है, और पैसे के लेन देन भी अंतरजाल पर होते हैं, इसीलिये यह आवश्यक है कि स्पैम को पहचाना जा सके और अंतरजाल को सुरक्षित बनाया जा सके। इसके लिए ये सभी तकनीकें बहुत वांछनीय हैं।

प्रयुक्त शब्दावली

Community Based Features	समुदाय आधारित विशेषताएं
Correlation	सह – संबंध
Cyber Crime	साइबर अपराध, अंतराजाल पर होने वाले अपराध
DR (Detection rate)	गलत सन्देश पकड़ने की दर
FN (False Negative)	जो सन्देश गलत था उसकी सही के रूप में गलती से हुई पहचान
FP (False Positive)	जो सन्देश गलत नहीं था उसकी गलत के रूप में गलती से हुई पहचान
FPR (False Positive Rate)	झूठी सकारात्मक दर
F Score	सटीकता / नीर क्षीर विवेकशीलता की शक्ति
Genuine (human) users	वास्तविक (मानव) उपयोगकर्ता

Metadata Based Techniques	पोस्ट या ट्वीट की सामग्री आधारित निर्णय व्यवस्था
OSN/ Open Social Network	अंतर्जाल पर उपलब्ध खुली हुई सामाजिक बातचीत आदि की सेवाएं जिन्हे कोई भी उपयोग में ले सकता है।
Phishing	व्यक्तिगत जानकारी, जैसे पासवर्ड और क्रेडिट कार्ड नंबर प्रकट करने के लिए व्यक्तियों को प्रेरित करने के लिए प्रतिष्ठित कंपनियों से ईमेल भेजने का फर्जी अभ्यास।
Random Forest, Decision Tree and Bayesian Network	वर्गीकरण / तुलना के तरीके
Random number Generator	अनियमित संख्याएँ बनाने वाले कम्प्यूटर प्रोग्राम
Real Time	जब अंतर्जाल पर जो हो रहा है उस पर आधारित त्वरित निर्णय उसी समय लिए जाना
Registered Follower	पंजीकृत अनुयायी
Social Bots	सामाजिक अंतरजाल पर मनुष्य की ही तरह झूठे पोस्ट या ट्वीट करने वाले सॉफ्टवेयर प्रोग्राम
Spam	लोगों को गुमराह करने या फंसाने के लिए भेजे गए झूठे सन्देश या पोस्ट
Spammer	झूठे संदेश या पोस्ट भेजने वाले स्रोत (मनुष्य भी और बॉट भी)
Time Activation Function	समय सक्रियण सबरूटीन जिससे ट्वीट ट्रिगर हो जाता है
TN (True negative)	जो सन्देश सही था (वास्तविक उपयोगकर्ता का सच्चा सन्देश) उसकी सही के रूप में सही पहचान
TP (True positive)	जो सन्देश गलत था उसकी गलत के रूप में सटीक पहचान

संदर्भ References :

1. Fazil, Mohd & Abulaish, Muhammad. (2018). “A Hybrid Approach for Detecting Automated Spammers in Twitter”. IEEE Transactions on Information Forensics and Security. PP. 1-1. 10.1109/TIFS.2018.2825958.
2. K. Soundararajan , U Eranna , Shilpa Mehta, “A Neural Technique for Classification of Intercepted e-mail Communications with Multilayer Perceptron using BPA with LMS Learning” International Journal of Advances in Electrical and Electronics Engineering, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.639.3947&rep=rep1&type=pdf
3. Shilpa Mehta , U Eranna , K. Soundararajan “A Fuzzy Technique for Classification of Intercepted Communication” International Journal of Communication Engineering Applications-IJCEA, ISSN: 2230-8520; e-ISSN: 2230-8539, pp 412- 416, Volume 3 Issue 1 https://www.researchgate.net/profile/Shilpa_Mehta2/publication/267558855_A_Fuzzy_Technique_for_Classification_of_Intercepted_Communication/links/578f372b08ae9754b7ecc2e1.pdf
4. M. Tsikerdekis, “Identity deception prevention using common contribution network data,” IEEE Transactions on Information Forensics and Security, 2017, vol. 12, no. 1, pp. 188–199.
5. S. Lee and J. Kim, “Warningbird: A near real-time detection system for suspicious urls in twitter stream,” IEEE Transaction on Dependable and Secure Computing, 2013, vol. 10, no. 3, pp. 183–195.
6. A. H. Wang, “Don’t follow me: Spam detection in twitter,” in Proc. SECRYPT, Athens, 2010, pp. 1– 10
7. Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, “Design and analysis of social botnet,” Computer Networks, 2013, vol. 57, no. 2, pp. 556– 578.
8. G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in Proc. ACSAC, Austin, Texas, 2010, pp. 1–9.
9. H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, “Sybilguard: Defending against sybil attacks via social networks,” IEEE/ACM Transactions on Networking, 2008, vol. 16, no. 3, pp. 576–589.