

USER SIMILARITY ON TWITTER: A DUAL PERSPECTIVE OF LITERATURE REVIEW AND EXPERIMENTAL COMPARISON

**A Thesis Submitted
In Partial Fulfillment of the Requirements for the
Degree of**

MASTERS OF TECHNOLOGY

in

Data Science

by

Vidushi Jain

(2K23/DSC/09)

Under the Supervision of

Dr. Sonika Dahiya

Assistant Professor, Department of Software Engineering

Delhi Technological University



**DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

June, 2025

**DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042**

CANDIDATE’S DECLARATION

I, Vidushi Jain, 2K23/DSC/09 students of M.Tech (Data Science), hereby certify that the work which is being presented in the thesis entitled “User Similarity on Twitter: A Dual Perspective of Literature Review and Experimental Comparison” in partial fulfilment of the requirements for the award of degree of Master of Technology, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from Jan 2025 to May 2025 under the supervision of Dr. Sonika Dahiya.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate’s Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s)

Signature of External Examiner

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “User Similarity on Twitter: A Dual Perspective of Literature Review and Experimental Comparison” which is submitted by Vidushi Jain, Roll No – 2K23/DSC/09, Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Dr. Sonika Dahiya

Assistant Professor

Date: 24.06.2025

Department of Software Engineering, DTU

**DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042**

ACKNOWLEDGEMENT

We wish to express our sincerest gratitude to Dr. Sonika Dahiya for her continuous guidance and mentorship that she provided us during the project. She showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. She was always ready to help us and clear our doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi

Vidushi Jain

Date: 24.06.2025

(2K23/DSC/09)

ABSTRACT

The rapid proliferation of social media platforms, particularly Twitter, has necessitated advanced techniques for understanding user behavior, identifying similar users, and uncovering community structures. This thesis presents a comprehensive study of methods for detecting user similarity and communities on Twitter, encompassing both literature review and comparative analysis perspectives.

The first part of the research synthesizes existing approaches into three primary categories: signal-based, machine-learning-based, and graph-based methods. These approaches leverage interaction patterns, social graph structures, and content alignment to address applications in security, audience targeting, and social recommendation. The strengths, limitations, and scalability of these methods are critically evaluated, with an emphasis on their adaptability to real-world scenarios and societal implications.

The second part focuses on a detailed comparative analysis of three established user similarity frameworks: TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users. Utilizing a dataset derived from the Twitter API, the study implemented ten similarity signals encompassing interaction, content, and network-based metrics. The results highlight strong correlations between interaction and retweet similarity metrics while underscoring the complementary insights of profile-based features. The computed rankings, derived from an aggregated similarity score, achieved a high Spearman correlation of 0.91 with human evaluations, validating the model's effectiveness.

This thesis concludes by identifying limitations and proposing future directions, such as adaptive weighting strategies, integration of temporal dynamics, and scalability testing for large datasets. By bridging theoretical insights with practical applications, this work contributes to the development of robust, adaptive, and interpretable systems for similarity detection and community discovery, enhancing the personalization and utility of social media platforms.

TABLE OF CONTENT

CANDIDATE’S DECLARATION	ii
CERTIFICATE	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
TABLE OF CONTENT	vi
LIST OF FIGURES.....	viii
LIST OF TABLES.....	ix
Chapter 1	1
Introduction	1
1.1 Problem Statement	1
1.2 Significance of User Similarity Detection	1
1.3 Motivation	2
1.4 Overview of Methods for Twitter User Similarity Detection	3
1.4.1 Interaction-Based Methods	3
1.4.2 Content-Based Methods	3
1.4.3 Graph-Based Methods.....	4
1.4.4 Hybrid Methods	4
1.5 Overview of Research Objectives	5
Chapter 2	7
Related Work.....	7
2.1 Overview of Existing Methods	7
2.2 Datasets and Data Collection	8
2.3 Methodologies and Techniques.....	9
2.3.1 Graph-Based Approaches.....	9
2.3.2 Machine learning (ML) based approaches	10
2.3.3 Signals-Based Methods.....	11
2.4 Evaluation Metrics and Analysis	12
2.5 Research Gaps	13
Chapter 3	14
Research Methodology.....	14
3.1 Framework Overview and Implementation Flow	14
3.2 Dataset Profile	15
3.3 Data Preprocessing	16
3.3.1 Data Cleaning.....	16
3.3.2 Text Processing	16
3.3.3 Feature Extraction	17
3.4 Techniques	17

3.4.1 TSim Framework	17
3.4.2 Characterizing and Detecting Similar Twitter Users	18
3.4.3 Self-Similarity of Twitter Users.....	19
3.5 Signal Computation.....	19
Chapter 4	21
Results And Discussion	21
4.1 Signal Score Analysis	21
4.1.1 Analysis of Interaction-Based Signals (S2, S8)	21
4.1.2 Analysis of Content-Based Signals (S3, S4, S5, S10)	22
4.1.3 Analysis of Graph-Based Signals (S1, S6, S7)	22
4.1.4 Analysis of Profile-Based Signal (S9)	23
4.2 Analysis and Visualization.....	23
4.2.1 Correlation Analysis of Similarity Measures	23
4.2.2 Candidate Ranking Process.....	24
4.3 Evaluation	25
4.3.1 Evaluation Approach.....	26
4.3.2 Evaluation Analysis	26
4.3.3 Reflection on Subjectivity	27
4.4 Discussion	27
Chapter 5	29
Conclusion and Future Scope	29
Bibliography	31

LIST OF FIGURES

2.1 Frequency of Papers Using Various Evaluation Metrics.....	13
3.1 Model Framework.....	14
4.1 Correlation Matrix of Similarity Measures.....	24

LIST OF TABLES

2.1 Datasets Used in Various Studies: Key Features and Annotation Status.....	8
2.2 Encapsulates graph-based approaches, their strengths and weaknesses	10
2.3 Encapsulates ML-based approaches, their strengths and weaknesses	10
2.4 Encapsulates signal-based approaches, their strengths and weaknesses	11
2.5 Summarizes the evaluation metrics used in each study and their results	12
3.1 Summarizes the evaluation metrics used in each study and their results.....	19
4.1 Similarity scores for all 11 candidate users (CU) across the ten signal types (S1-S	21
4.2 Candidates ranked based on composite score	25
4.3 Candidates computed rank versus human-evaluated rank	27

CHAPTER 1

INTRODUCTION

With changing social media patterns at a rapid pace and the increasing power of online interactions, discovering similar Twitter users is gaining prominence as an important task for recommendation targeting, community analysis, and influence assessment [1]. The subsequent problem statement defines the prominent challenges and research gaps to be addressed by this research in detecting similarity among users using interaction, content, and graph signals.

1.1 Problem Statement

Identification of similar Twitter users has become a very significant area in the context of applications like targeted recommendation, influencer detection, and community analysis [1], [2]. Current literature presents varying methods for detecting user similarity that include signal-based, graph-based, and machine learning approaches, having varying advantages and limitations. But most of the current research considers disconnected user behavior aspects, e.g., interaction behaviors, overlap of content, or network structure, without thoroughly fusing these aspects.

To fill these gaps, this thesis will implement and compare the performance of three existing frameworks—Tsim [1], Characterizing and Detecting Similar Twitter Users [2], and Self-Similarity of Twitter Users [3]—in identifying user similarity based on interaction, content, and graph-based signals. By integrating results of the literature review and applying the three frameworks, the research aims to measure the predictive capacity of individual signals, investigate their interconnection, and suggest a composite similarity model that weighs different types of signals. The research also tests the scalability and reliability of the suggested model, offering knowledge on its usability in large Twitter datasets.

1.2 Significance of User Similarity Detection

Finding similar social media users of the same class on networks such as Twitter is crucial for achieving the optimum level of user interaction, content recommendations, and community research. By finding users with the same kind of interaction patterns, similar interests, or similar social networks, networks can use personalization algorithms, recommending users similar content, accounts, and topics to them. Not only does this optimize user satisfaction but also helps in site sticking by showing more engaging online interactions [1], [2].

Other than recommendation systems, user proximity detection is also critical in social network analysis and community detection. Clustering adjacent users uncovers thematic communities and social structures and therefore helps in the creation

of targeted communication campaigns. For example, for influencer marketing, brands can target similar-interest groups effectively and thereby maximize campaign reach and influence [4].

Also, the capacity to recognize analogous users has wider social applications, particularly for health promotion in the public sphere and crisis communication. From the analysis of the users posting similar content or exhibiting similar behavior patterns, public authorities can deploy targeted messages during health crises or natural disasters and thereby increase the reach and coverage of imperative information [4]. Similarly, identifying user similarity is vital in the detection of coordinated disinformation campaigns, allowing platforms to proactively counteract the dissemination of spurious information.

From the research point of view, the combination of interaction, content, and graph-based signals provides a richer picture of user relationships, overcoming the shortcomings of current approaches that are based only on isolated types of signals. This thesis makes use of the frameworks of TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users for testing the validity of multi-signal methods with a view to creating a stronger and more understandable similarity model that balances diverse signal types well.

1.3 Motivation

Although user similarity discovery is important, an extensive literature review on this matter was hitherto missing, creating a void for synthesizing current methods and determining trends in research. Filling this void, the current research initiated a systematic review of eight pioneer papers, classifying current methods into interaction-based, content-based, and graph-based signals.

In addition, while current methods tend to concentrate on singular signal types—e.g., retweet behavior, hashtag overlap, or follower relations—they do not capture the entire range of user similarity. Aware of this constraint, this research endeavors to apply and contrast three proven frameworks—TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users—each capturing unique signal combinations. The objective is to assess the ability of solitary signals to predict, investigate their relationship, and propose a more comprehensive similarity model with interaction, content, and graph-based features.

By bridging the gap in the literature and providing a unified similarity detection model, this work contributes to the promotion of recommendation systems, influencer research, and community discovery on Twitter, enhancing more efficient user interaction and content targeting.

1.4 Overview of Methods for Twitter User Similarity Detection

The current methods of detecting user similarity can be generally divided into three categories: interaction-based, content-based, and graph-based methods.

1.4.1 Interaction-Based Methods

Interaction-based approaches evaluate user similarity based on explicit user interactions, including mentions, retweets, replies, and likes [1], [3]. Such approaches utilize the generalization that users who interact most with each other or have similar interactions tend to be similar.

- **Retweet Similarity:** An approach that compares users' retweet patterns to find common content interest.
- **Mention Similarity:** Calculates the degree of direct interaction in terms of mentions and replies.
- **Favorite Similarity:** Computes similarity between users on the basis of liked content.
- **Interaction Frequency:** Measures interaction intensity, quoting rates of user behavior.
- **Limitations:** Interaction-based techniques are neglectful of users with common content interest but no direct interaction. They are also time-variant as trends in interaction vary over time.

1.4.2 Content-Based Methods

Content-based approaches emphasize content examination of user-generated information, for example, hashtags, topics, and user-generated tweets [1], [2]. These approaches struggle to ascertain content alignment through exploration of thematic similarity among users.

- **Hashtag Similarity:** Tracks hashtag overlap to quantify content alignment.
- **Topic Modeling:** Finds latent topics using methods like Latent Dirichlet Allocation (LDA) to determine similar topics.
- **Textual Similarity:** Quantifies text in tweets using methods like cosine similarity or Jaccard index.
- **Profile Similarity:** Computes profile properties like bio, age and interests to deduce user similarity.

- Limitations: None of the content-based approaches perform well in obtaining strong social relationships and are very sensitive to text quality and topic model results.

1.4.3 Graph-Based Methods

Graph-based methods use the social network structure, e.g., follow relationships, common friends, and network clusters, to determine similar users [1], [2]. They are highly effective in discovering community structures and influential users.

- Followings and Followers Similarity: Identifies common friends within the social graph.
- Common Friends Analysis: Approximates the number of shared friends to quantify network proximity.
- Network Clustering: Applies graph clustering algorithms to discover groups of similar users.
- Centrality Measures: Quantify influential user location in the network, i.e., degree centrality or PageRank.
- Limitations: Graph-based methods are computationally expensive on large data and can possibly overlook content-based similarities among users.

1.4.4 Hybrid Methods

Some frameworks combine interaction, content, and graph-based methods to achieve a more comprehensive similarity measure. The three frameworks used in this study—TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users—are some of the multi-dimensional methods:

1. TSim Framework (Hind AlMahmoud) [1] : TSim integrates seven signals such as interaction (retweets, mentions, favorites), content (shared interests, hashtags), and profile attributes (language, bio). It is a MapReduce distributed programming paradigm that can handle large data sets and hence scale. Through the combination of the different signals, TSim has a well-balanced user similarity measure and is highly efficient in identifying users with shared interests and frequent interactions.
2. Characterizing and Detecting Similar Twitter Users (Ali Choumane) [2]: This model relies on social graph and content similarity using three main signals - Followings Signal that detects direct follow relations, common friends signal that computes the intersection of mutual friends, capturing network proximity, Top-10 Topics Signal that derives top 10 topics by applying LDA and computes

them by cosine similarity to measure thematic similarity. By integrating social network topology and content topics, this model effectively detects thematic communities of users that are alike.

3. Self-Similarity of Twitter Users (Masoud Fatemi) [3]: This model emphasizes patterns of interaction and behavior profiles, utilizing signals like Interaction-Based Similarity which examines mentions, retweets, and replies to calculate engagement, Profile-Based Similarity which Compares user behavior in terms of attributes like tweet frequency, favorites, and account age. Hashtag-Based Similarity which measures content alignment by common usage of hashtags. The emphasis on profile information and interaction data makes this model highly effective for the identification of users who have similar patterns of interaction, even without direct social relationships.

1.5 Overview of Research Objectives

The main aim of this dissertation is to compare and assess the efficacy of interaction, content, and graph-based signals in identifying Twitter user similarity based on three well-established frameworks: TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users. Through the application and exploration of these frameworks, this study sets out to meet the following specific research aims:

- To undertake an in-depth literature review of the current Twitter user similarity detection techniques, classifying them into interaction-based, content-based, graph-based, and hybrid techniques, and highlighting important limitations and research gaps.
- To implement three selected frameworks (TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users) and compute similarity signals for a collection of Twitter users based on interaction, content, and graph-based measures.
- To produce a composite similarity model that aggregates several signals using a weighted scoring mechanism to provide a total measure of user similarity across interaction, content, and network structure.
- To perform correlation analysis and consistency checking of the computed similarity signals to find interrelations among signals and their individual and collective predictive ability to identify similar users.
- To compare the composite similarity model with human-rankings of similarity, to test the effectiveness and robustness of the model using statistical measures such as Spearman correlation.
- To contrast the scalability and practicability of the proposed similarity model

by measuring performance across different sets of users and approximating its possible pitfalls in large-scale applications.

- To provide recommendations and suggestions for future research emphasizing major signal types that make the largest contributions in capturing similarity between users and offering direction for the inclusion of advanced techniques such as temporal analysis and machine learning.

CHAPTER 2

RELATED WORK

The identification of similar users on Twitter has emerged as a significant research area with applications in recommendation systems, community detection, and content analysis. Numerous studies have explored interaction-based, content-based, and graph-based approaches to assess user similarity. This chapter systematically reviews the existing literature, focusing on the datasets used, methodologies employed, and evaluation metrics to highlight the strengths and limitations of prominent frameworks in the domain and finally suggest the research gaps as found existing methods.

2.1 Overview of Existing Methods

Hind AlMahmoud [1], developed a system named TSim, which measures the user's similarity based on seven different signals which includes followings and followers, mention, retweet, favorite, common hashtag, common interests, and profile similarity. The framework is implemented using the MapReduce distributed programming model, which allows for the scalable processing of large datasets. This ensures the system can handle big data typical of social media platforms like Twitter.

Andrea Tundis [5], focused on a particular application (analyzing criminals) of user similarity instead of generalizing it. It proposes a method using text analysis and metadata to analyze criminals on social media, identify criminal activities. It uses cosine and jaccard similarity to compute the similarity score. And for text analysis, they used TF-IDF and Latent Dirichlet Allocation (LDA) to identify important topics shared by users. Finally clustering techniques are applied to group users based on their interests and behavior.

Shaohua Tao [6], proposed a graph based method to calculate similarity between two users and recommend them as friend to source user. It builds a knowledge graph from the wikipedia documents and calculates the shortest distance between the concepts belonging to different user using the WESP (Weighted Euclidean Shortest Path) method. The shorter the path, the more similar two users are.

Ali Choumane [2], proposed three signals based on social graph and user profile contents. The three signals include the following signal, top 10 topics signal and common friends signals. These signals were statistically compared with existing signals and several classifiers were built on real Twitter data which in result proved the better performance of proposed signals compared to existing ones.

In Masoud Fatemi [3], the study employs methods to measure similarity between users based on interactions between a user (ego) and other users (alters), profile-based activity history and linguistic content of tweets.

Niloufar Shoeibi [7], proposed hybrid model provides a sophisticated and effective way to measure Twitter profile similarity by integrating behavioral, network, and content-based features. They have used Dynamic Time Warping (DTW)

to measure similarity between behavioral ratios, Jaccard similarity for measuring similarity between audience, cosine similarity for measuring similarity between content.

Md Ahsan Ul Hasan [4], proposed a Twitter community detection method that evaluates similarity based on user profiles, tweet topics, and tweet sentiment. They built a similarity network and compared it to Zachary's Karate Club Network. Their approach improves the modularity and conductance quality functions in the Louvain community detection algorithm, leading to a more accurate and insightful community structure within the Twitter network.

Siyi Guo [8], proposed SoMeR, a Social Media user Representation learning framework that incorporates temporal activities, text content, profile information, and network interactions to learn comprehensive user portraits. SoMeR encodes user post streams as sequences of time stamped textual features, uses transformers to embed this along with profile data, and jointly trains with link prediction and contrastive learning objectives to capture user similarity.

This systematic review aims to summarize the various approaches used to measure user similarity on Twitter and to explore the advantages and limitations of these methods. This review includes only post-2016 publications.

2.2 Datasets and Data Collection

This section provides the details of the data sources, highlights the main features extracted for model development, and indicates whether annotation was performed. Table 2.1 summarizes the datasets used in key studies to measure similarity among Twitter users.

Table 2.1. Datasets Used in Various Studies: Key Features and Annotation Status

S.No.	Reference	Data Used	Main features of data	Origin	Annotations done
1	Hind AlMahmoud [1]	Manually taking a bunch of Twitter users.	Following/Follower list, retweets, hashtag, profile.	Manual Selection	No
2	Andrea Tundis [5]	“AboutIsis” Kaggle dataset.	Profile data, friends/followers, user’s interest	Available on Kaggle [9]	Yes
3	Shaohua Tao [6]	Two hundred source seed users’ and their followers’ tweets were extracted	Tweet topics	Twitter API	Yes

		resulting in 200 million tweets.			
4	Ali Choumane [2]	The dataset is 3.5 GB and consists of 387 pairs of followers users	Tweets, friends,	Twitter API	Yes
5	Masoud Fatemi [3]	16,816,460 tweets were extracted from 8,744 accounts via the Twitter API.	Friends, tweets, retweets, mentions, quotations	Twitter API	No
6	Niloufar Shoeibi [7]	A balanced dataset of 19,900 entries was created, using data extracted via the Twitter API.	Tweets, Retweets, mentions, replies, statuses, likes	Twitter API	Yes
7	Md Ahsan Ul Hasan [4]	The study uses real-time data from the Twitter API [10]	User profiles, tweet subjects, and tweet sentiments.	Twitter API	Yes
8	Siyi Guo [8]	The study uses datasets from [14], and [15].	Hashtags, followings, and follower, retweets	Available on GitHub [13]	No

All the above-mentioned datasets are accessible. However, to access Twitter API features, we need some level of additional access.

2.3 Methodologies and Techniques

This section outlines the methodology for assessing similarity between Twitter users, utilizing various approaches. These methods include graph-based techniques, clustering techniques, and signals derived from follower-following relationships, retweets, hashtags and topics.

2.3.1 Graph-Based Approaches

In graph-based approaches, the Twitter network is modeled as a graph structure where various elements such as users, tweets, hashtags, and retweets are represented as nodes. The interactions between these elements (like follows, mentions, or retweets) are represented as edges. This graph structure is then analyzed using graph algorithms to calculate user similarity. Shaohua Tao's [6], and Md Ahsan Ul Hasan's [4] studies use a graph-based approach to measure user similarity. Table 2.2 encapsulates the graph-based approaches.

Table 2.2. Encapsulates graph-based approaches, their strengths and weaknesses

S. No.	Reference	Methodology	Strengths	Weaknesses
1	Shaohua Tao [6]	The Weighted Euclidean-Shortest Path (WESP) method was introduced, with an optimized similarity measurement (OSM) model enhancing its efficiency.	The OSM model outperforms the baseline methods.	It may require more computational resources.
2	Md Ahsan Ul Hasan [4]	Louvain community detection algorithm [14] is used to identify optimal group structures in the Neo4j similarity network	Neo4j is effectively utilized to model complex relationships.	The study does not explore other community detection algorithms.

2.3.2 Machine learning (ML) based approaches

ML-based approaches leverage clustering and classification algorithms to identify and group similar Twitter users. By extracting a set of meaningful features from user profiles and interactions, these techniques can uncover patterns that define user similarity and categorize users accordingly. Andrea Tundis's [5], and Siyi Guo's [8] studies use a machine learning-based approach to measure user similarity. Table 2.3 encapsulates the ML-based approaches.

Table 2.3. Encapsulates ML-based approaches, their strengths and weaknesses

S. No.	Reference	Methodology	Strengths	Weaknesses
1	Andrea Tundis [5]	Used Jaccard and Cosine similarity for profile similarity, LDA and TF-IDF for text analysis, and clustering techniques for behavior similarity. Then used random forest classifier.	The introduction of the Corrective Factor (CF) to manage missing values in user data is a significant contribution.	A need to expand to include more complex content like images and videos.

2	Siyi Guo [8]	Used a transformer model to create user representations from social media data, predicting network links and grouping users with similar behaviors.	It leverages temporal features to monitor changes in user behavior, offering insights into evolving trends.	Underutilization of Temporal Features. It could benefit from more advanced temporal analysis.
3	Niloufar Shoeibi [7]	Used Dynamic Time Warping [15] for behavior similarity, Jaccard similarity for audience Network similarity, cosine similarity for content similarity, and train random-forest classifier.	The methodology combines multiple dimensions of profile similarity leading to a robust similarity measurement.	The proposed method's applicability requires future improvements such as gender analysis.

2.3.3 Signals based approaches

The signals-based approach introduces specific metrics or formulas, known as signals, to quantify the similarity between Twitter users. These signals are derived from user interactions and behaviors on the platform. By aggregating these signals, we can compute a similarity score that reflects how alike two users are. Hind AlMahmoud's [1], Ali Choumane's [2], Masoud Fatemi's [3], Niloufar Shoeibi's [7] studies use a signal-based approach to measure user similarity. Table 2.4 encapsulates the signal-based approaches.

Table 2.4. Encapsulates signal-based approaches, their strengths and weaknesses

S. No.	Reference	Methodology	Strengths	Weaknesses
1	Hind AlMahmoud [1]	Proposed TSim which uses seven signals and implemented using the MapReduce model.	Scalable, flexible, and a comprehensive similarity formula.	Customizable weights may lead to users overfitting the model.
2	Ali Choumane [2]	Built classifiers like Multilayer Perceptron, SVM, Naïve Bayes, and k-Nearest Neighbors using the proposed signals.	Outperforms existing methods in precision for detecting similar users.	A significant challenge of data labeling, affecting the accuracy of classifiers.
3	Masoud	Ranked users to	It reveals that	The study reveals

Fatemi [3]	measure similarity using interaction-based, profile-based, and hashtag-based similarity formulas.	user interactions are a more potent indicator of similarity.	that larger networks reduce the accuracy of similarity detection.
------------	---	--	---

2.4 Evaluation Metrics and Analysis

This section discusses the key evaluation metrics used in various proposed approaches for Twitter user similarity detection. Table 2.5. summarizes the evaluation metrics used in each study and their results and Fig 2.1. depicts the frequency of paper using various evaluation metrics.

Table 2.5. Summarizes the evaluation metrics used in each study and their results

S. No.	Reference	Evaluation Metrics	Results
1	Hind AlMahmoud [1]	Human Judges, Twitter's own Who To Follow (WTF) [16]	The results were promising and reasonably accurate.
2	Ali Choumane [2]	Precision	Proposed signals performed superior than SVO Hashtags and TSim in every classifier.
3	Masoud Fatemi [3]	Accuracy, Pearson correlation	On average, interaction-based similarity is the most accurate with 19.2% accuracy of the three proposed signals. The size of ego networks and the suggested approach have a negative linear correlation.
4	Niloufar Shoeibi [7]	Accuracy	Random forest classification performed the best with 97% accuracy.
5	Siyi Guo [8]	ROC-AUC, F1 Score	The proposed method outperforms the baselines.
6	Md Ahsan Ul Hasan [4]	Compared with Baseline Zachary's Karate Network	The proposed method achieves a higher modularity score (0.5269) compared to the benchmark (0.4616).
7	Andrea Tundis [5]	Accuracy, precision, recall	Achieved accuracy and precision equal to 92%, and a recall of 100%.

8	SHAOHUA TAO [6]	Spearman correlation coefficient, Precision	OSM outperformed the baseline methods.
---	--------------------	--	---

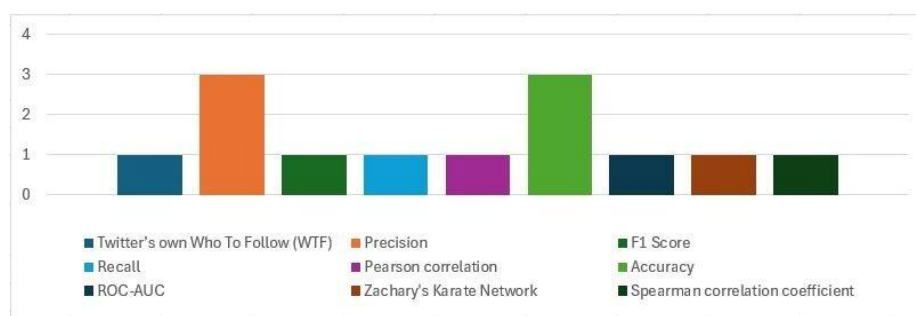


Fig 2.1. Frequency of Papers Using Various Evaluation Metrics

2.5 Research Gaps

Even with significant research on Twitter user similarity detection, some areas are left untouched:

- **Integration of Multiple Signal Types:** All current approaches are centered around one type of signal — interaction-based, content-based, or graph-based — and do not synthesize them effectively to find a combined similarity measure. TSim synthesizes diverse signals but fails to conduct an in-depth intensity analysis of interactions. Characterizing and Detecting Similar Twitter Users successfully captures content and graph signals but fails to identify interaction patterns. Filling this knowledge gap in this thesis by using a composite model that synthesizes all three aspects could improve the accuracy of similarity detection.
- **Scalability Challenges:** Computation-intensive approaches like LDA and cosine similarity are beset with scalability challenges in handling big data. Although TSim addresses this with the use of MapReduce, there are other systems that lack scalability mechanisms naturally, particularly for interactive real-time analysis.
- **Temporal Analysis:** The majority of the models depend on static data and fail to consider changing patterns of user behavior over time. Temporal analysis can enable more context-aware measures of similarity to identify changing topics of content as well as changing interaction strengths.
- **Evaluation Metrics:** A majority of the current frameworks use quantitative measures (e.g., cosine similarity, Euclidean distance) with little human-evaluated validation. Employing labelled datasets or expert judgement can be utilized to enhance comparison of similarity scores.

CHAPTER 3

RESEARCH METHODOLOGY

The approach employed in this study is to organize an analysis of the effectiveness of content-based, interaction-based, and graph-based signals in determining Twitter user similarity in three established frameworks: TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users. The approach has the following key elements.

3.1 Framework Overview and Implementation Flow

The architecture suggested combines interaction-based, content-based, and graph-based signals for extensive evaluation of user similarity on Twitter using three existing architectures: TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users. Procedures for data collection involve retrieving interaction data (mentions, retweets, favorites), content data (hashtags, tweet content), and network data (followings, followers) of 1 tested user and 11 candidate users. Preprocessing is done for cleaning and normalization of the data so that there can be homogeneity for computation of signal.

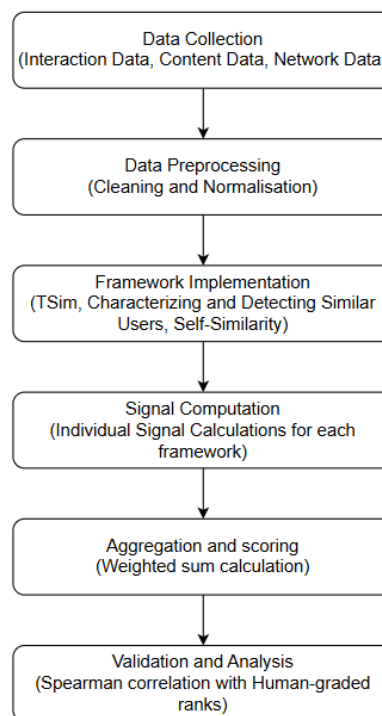


Fig 3.1 Model Framework

Every model is run independently using independent sets of signals in order to compute the similarity of the users. TSim employs seven signals such as user interactions, content similarity, and profile characteristics, while social graph structure and thematic analysis of the content are employed by the Characterizing and Detecting Similar Twitter Users model. The Self-Similarity approach focuses on interaction intensity and profile-based behavior to determine co-engaged users with similar engagement patterns. The computed signals are then aggregated using a weighted sum technique to provide a final similarity score, which is thereafter validated against human-graded similarity rankings using Spearman correlation.

3.2 Dataset Profile

Data for this analysis was gathered through Twitter API, in compliance with privacy guidelines by anonymizing the names of profiles and giving each user a unique identifier (X1–X11). The dataset included 1 studied user and 11 candidate users, who were chosen by their interaction pattern, content features, and network structure.

The data collection process entailed the retrieval of three classes of important data categories, which included profile data, tweet texts, and social network topology. The major constituents of the data collection are listed below:

1. Profile Data:

- For every user, core profile metadata like account creation time, bio, location, and language were obtained.
- It was utilized to compute profile similarity so that users with similar profile attributes could be discovered.

2. Tweet Content:

- Around 500 tweets per user were collected, both retweets, mentions, and original posts.
- Text content was preprocessed to identify: Hashtags - Extracted for determining similarity of content based on mutual application of hashtags, Mentions – Verified to assess interaction-based similarity, Sentiment Analysis – Performed to identify user sentiment in common topics or hashtags.

3. Interaction Data:

- In order to determine interaction-based similarity, retweets of candidate user tweets by the user under study were counted.
- This provided insights into shared content engagement, indicating potential content alignment between users.

4. Social Network Structure:

- The followers and followings lists of each user were retrieved, capturing the complete social network structure.
- Since all the users had less than 5000 connections, their entire network graph could be downloaded and analyzed.
- Network data was used to compute signals such as:

- Followings and Followers Similarity – Measuring network overlap between users.
- Common Friends Signal – Identifying shared connections to infer social proximity.

5. Data Anonymization and Ethical Considerations:

- Anonymity of the users was ensured by providing IDs (X1–X11) to avoid their individual information being disclosed.
- The data collection strictly followed Twitter data access policies for upholding users' privacy and ethical use of data.

This integrated data collection architecture enabled the extraction of different types of signal on interaction, content, and graph-based dimensions to build upon in later signal computation and similarity analysis within the research.

3.3 Data Preprocessing

Data preprocessing is an important step to guarantee the uniformity, reliability, and analytical correctness of the collected Twitter data used in this research. The data gathered consists of tweets, user profile data, and network structure data for 1 analyzed user and 11 candidate users with up to 500 tweets each. This section details the individual preprocessing processes that were implemented on the dataset, grouped under data cleaning, text processing, feature extraction, and data transformation.

3.3.1 Data Cleaning

Data cleaning refers to the elimination of unnecessary, redundant, or missing data to guarantee the quality of the dataset. The steps employed were:

- **Filtering Non-English Tweets:** As the study is centered on textual content analysis, only tweets in English (`lang = "en"`) were kept. Non-English tweets were removed to ensure language uniformity.
- **Time Formatting:** The `created_at` field, initially in string format (e.g., "Fri Sep 10 10:15:32 +0000 2025"), was normalized to a uniform datetime format (YYYY-MM-DD HH:MM:SS) to be treated uniformly.

3.3.2 Text Processing

Tweet text data would likely be noisy and unstructured containing URLs, non-standard words, and special characters. The following preprocessing steps were undertaken:

- **Lowercasing:** All content in the `full_text` field was converted to lower case to eliminate inconsistencies because of capitalization.
- **Removing URLs and Emojis:** URLs and emojis, which are not of value for content analysis, were removed using regex patterns.

- **Tokenization and Stopword Removal:** Text content was tokenized to words and frequent stop words (e.g., "and", "the", "of") eliminated using the NLTK library
- **Lemmatization and Stemming:**
 - To standardize textual content, words were reduced to their root forms using lemmatization (e.g., "running" → "run").
 - This ensures that similar words are treated as identical during content analysis and similarity computation.

3.3.3 Feature Extraction

From the processed data, specific features relevant to the similarity signals were extracted and structured for analysis. These include:

- **Hashtags and Mentions Extraction:**
 - Hashtags and user mentions were extracted separately to facilitate content and interaction analysis.
 - Extracted hashtags were kept in list format as a comma-separated list, whereas mentions were counted as a single feature.
- **Network Structure Extraction:**
 - The followers and followings list for each user was retrieved in order to construct the network graph.
 - The entire set of followings of followers and followers of followings were also retrieved in order to aid graph-based similarity computation.

3.4 Techniques

In this study, three well-known Twitter user similarity detection frameworks are applied: TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users. Each of them uses different sets of signals that cover interaction, content, and graph-based approaches in order to thoroughly examine user similarity. The frameworks are explained as follows:

3.4.1 TSim Framework

TSim framework, by Hind AlMahmoud [1], calculates user similarity based on a combination of seven prominent signals, each indicating a particular facet of user behavior. The framework supports large data sets through the MapReduce programming model, thereby supporting scalability and efficient processing. The seven signals employed in TSim are:

- **Followings-Followers Similarity:** This graph-based signal calculates the overlap in the number of followers and followings between two users. The

similarity value is normalized by the sum of connections.

- **Retweet Similarity:** This interaction feature calculates to what extent the subject user retweets the candidate user's tweets. The measure is normalized by the sum of retweets.
- **Mention Similarity:** Approximates the intensity of interaction by the number of times the subject user mentions the candidate user.
- **Favourite Similarity:** Compares content closeness based on the subject user favoriting the candidate user's set of tweets.
- **Hashtag Similarity:** This signal based on content computes the intersection over union of hashtags associated with the subject and candidate users to extract common thematic interests.
- **Common Interests:** Compares top five most frequent interest categories and identifies common themes, e.g., politics, sport, or technology. Topic extraction is conducted against a precompiled lexicon.
- **Profile Similarity:** Compares user similarity on the parameters of gender, language, and location, with a high score of 3 points, 1 point for every matched parameter.

The final similarity score in TSim is determined as the weighted sum of the component signals so that the model will be able to adjust the impact of each signal according to its relative significance to the specific application.

3.4.2 Characterizing and Detecting Similar Twitter Users

The method put forward by Ali Choumane highlights the combination of graph and content-based signals to enhance detection of user similarity [2]. It introduces three new signals which seek to capture the structure of the network and overlap of content:

- **Followings Signal:** This graph-based signal provides the proportion of the analyzed user's followings who also follow the candidate user. It succeeds in modeling social proximity in the network.
- **Common Friends Signal:** Counts the number of shared followings between the test user and candidate to identify common friends and community structure.
- **Top-10 Topics Signal:** This signal, based on content, identifies the top 10 topics discussed by every user using Latent Dirichlet Allocation (LDA) and then computes the cosine similarity in the topic distribution to measure thematic similarity.

By integrating network structure and thematic topics, the presented architecture precisely identifies users with common thematic interests and social connections.

3.4.3 Self-Similarity of Twitter Users

The proposed architecture by Masoud Fatemi takes into account user similarity based on interaction behavior, profile characteristics, and hashtags [3]. The approach best suits identifying users with similar engagement behavior even without explicit social relations. The three main signals are:

- **Interaction-Based Similarity:** This interaction-based metric measures the overall frequency of retweets and mentions to compute how much the candidate user is interacted with by the user under observation.
- **Profile-Based Similarity:** Each user is built a profile based on characteristics like account age, tweet frequency, and reputation score. Both profiles' Euclidean distance is computed, with lower distances reflecting greater similarity.
- **Hashtag-Based Similarity:** This signal based on content is the rate and pattern of hashtags employed by both sides and determining a similarity score as a function of their respective hashtag use patterns.

3.5 Signal Computation

From the gathered Twitter data, several features were extracted to realise the similarity signals as proposed in the above frameworks. Table 3.1 shows the most related signals, their category, and a succinct definition:

Table 3.1. Summarizes the evaluation metrics used in each study and their results

S.No.	Signal Name	Signal Type	Definition
S1	TSim Following-Follower [1]	Graph-based	Overlap in followers/followings, normalized by total connections.
S2	TSim Retweet [1]	Interaction-based	The count of retweets by the examined user of the candidate's tweets normalized.
S3	TSim Hashtag [1]	Content-based	The offsets quantify differences in sentiment-based tweet counts for each hashtag.
S4	Tsim Common Interests [1]	Content-based	Intersection of top 5 interests from a predefined lexicon (e.g., "Politics").
S5	Top 10 Topics [2]	Content-based	Cosine similarity of LDA-derived topic distributions [17] from tweets.

S6	Followings Signal [2]	Graph-based	Fraction of examined user's followings who follow the candidate.
S7	Common Friends [2]	Graph-based	Count of mutual followings of examined and candidate users.
S8	Interaction Similarity [3]	Interaction-based	The sum of mentions and retweets of candidate users.
S9	Profile Similarity [3]	Profile and content-based	Creates a profile for each Twitter user using the features: age, total tweets, reputation, tweet rate, hashtags, and hashtags density and calculates the Euclidean distance between their feature values.
S10	Hashtag Similarity [3]	Content-based	It compares the frequencies of hashtags used by two users. It calculates a "score" to measure how aligned their hashtag usage is.

CHAPTER 4

RESULTS AND DISCUSSION

This section includes the results of using interaction-based, content-based and graph-based methods to compare the examined and candidate Twitter users. To analyze the signals, we compute similarity averages for them, do a correlation analysis to find out how the signals relate to one another and check the results from both the algorithms and the human ratings. After that, the findings are reviewed to notice the main signals and assess if the suggested framework meets the goals.

4.1 Signal Score Analysis

In this section, we will see the score of similarity for all the signals among the 11 potential users. The purpose is to review all the kinds of signals used for user similarity detection, as well as key signals and determine how much the similarity score varies between user candidates. Table 4.1 below summarizes the similarity scores for all 11 candidate users across the ten signal types:

Table 4.1. Similarity scores for all 11 candidate users (CU) across the ten signal types (S1-S10)

CU	S1	S2	S3	S4	S ₅	S6	S7	S8	S9	S10
X1	9.75	29	3	3	1	0.13	0.55	56	1.01	0.31
X2	5.02	13	3	3	0	0.30	0.59	33	0.69	0.33
X3	6.33	0	2	4	0	0.15	0.31	3	1.84	0.05
X4	4.86	3	2	5	1	0.12	0.78	5	0.34	0.33
X5	5.37	2	1	3	0	0.16	0.32	3	1.43	0.07
X6	2.96	0	2	3	1	0.09	0.54	0	0.76	0.14
X7	1.33	0	1	3	0	0.00	0.00	0	1.13	0.02
X8	2.48	0	2	3	1	0.15	0.42	0	0.47	0.16
X9	7.39	0	1	3	0	0.10	0.44	0	1.00	0.14
X10	5.44	13	2	3	1	0.21	0.42	43	1.18	0.22
X11	0	0	0	1	0	0.00	0.00	0	0.34	0.00

4.1.1 Analysis of Interaction-Based Signals (S2, S8)

Interaction-based signals capture user engagement through **retweet and mention frequency**:

- **Retweet Similarity (S2):**
 - Retweeting similarity is highest for X1 with the examined user which is 29.
 - X2 and X10 have 13 retweet scores each, signaling that their content could be connected by retweets between the two.
 - Meanwhile, users who are X6, X7 and X9 do not retweet, suggesting that they share very little with others.
- **Interaction Similarity (S8):**
 - X1 and X10 both achieved very high scores (56 and 43), suggesting a strong bond with the analyzed user.
 - X6, X7, X8 and X9 do not interact with each other at all, further proving they do not have any meaningful exchange.

4.1.2 Analysis of Content-Based Signals (S3, S4, S5, S10)

These signals analyze users' behavior by looking at hashtags, what they are interested in and what subjects they like.

- **Hashtag Similarity (S3):**
 - X1, X2 and X2 are tied as the most likely users to discuss similar trending subjects, as their scores are 3, 3 and 2, respectively.
 - A little overlap between X7 and X9 indicates that the content they post is not very similar.
- **Common Interests (S4):**
 - The highest score of 5 on X4 means that there is a high degree of commonality in the interest categories.
 - Scores of 3 for X1, X2 and X3 suggest that all three have some similarities in their interests.
- **Top-10 Topics Signal (S5):**
 - The topics discussed by most candidates are not very similar, with top subjects showing a score of 1.
- **Hashtag Usage Pattern (S10):**
 - Most users still have a low hashtag alignment score, with both X1 and X2 having the highest values (0.31 and 0.33).

4.1.3 Analysis of Graph-Based Signals (S1, S6, S7)

Graph-based signals focus more on the structure and repeated connections in the neural network.

- **Followings-Followers Similarity (S1):**

- X1 gets a score of 9.75 which means there is a significant overlap between X1's and X2's networks.
- The score of 7.39 for X9 implies the network is moderately connected.
- X7 and X11 share amongst the lowest amount of connections.
- **Followings Signal (S6):**
 - The highest result on the signal score shows that X2 has many followers that are also followed by others (0.30).
- **Common Friends Signal (S7):**
 - X4 shows the greatest relationship in common (0.78).
 - It seems both X7 and X11 are isolated, since they have no mutual friends.

4.1.4 Profile-Based Signal (S9)

- **Profile Similarity (S9):**
 - The profile similarity score is based on the Euclidean distance between the profiles (e.g., how old the account is, how many tweets it posts).
 - X3 has a similarity score of 1.84 which means the two profiles are very alike.
 - In contrast, X4 and X11 have the least similarity in their profiles with a score of 0.34.

4.2 Analysis and Visualization

Here, we discuss in detail the techniques used for analysis and visualization in the study. It includes analysis of similarity measures, uses a heatmap to compare them and ranks users according to their similarity scores.

4.2.1 Correlation Analysis of Similarity Measures

We found out how the various similarity measures are connected by doing a correlation analysis. The aim was to pick indicators that have a clear positive link, as this could suggest that they are complementary. As a result of this analysis, you can understand which signals are similar and which bring something different to the evaluation process.

The key findings from the correlation analysis include:

- **Interaction and Retweet Similarity:** A strong positive correlation (0.96) was observed, indicating that the stronger the interaction between users and the examined user, the more likely it is that they will be retweeted by them. This points out the connection between people who engage and those who share information.
- **Hashtag Similarity and Common Friends:** Hashtag Similarity and Common

Friends show a strong correlation of 0.87. This means people who use the same hashtags are likely to have friends in common, suggesting that interests and social connections are strongly related.

- **Followings and Hashtag Similarity:** These measures showed a moderate positive correlation of 0.72, implying that users who have similar accounts they follow usually interact with similar-themed content on social media.

Interestingly, there were no significant negative correlations among the measures, suggesting that they generally complement one another. This indicates that the different signals work together to provide a more comprehensive evaluation of user similarity.

To make these relationships more accessible, a heatmap was created (Figure 1). With the heatmap, we can see which similarity measures have a strong connection to each other.

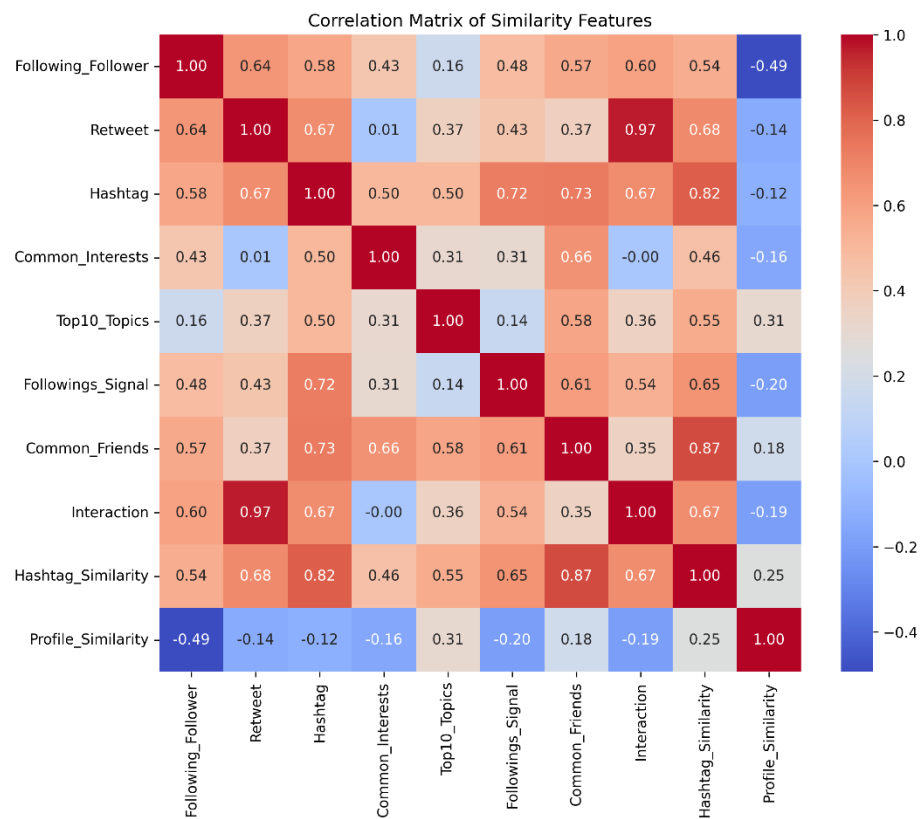


Fig 4.1 Correlation Matrix of Similarity Measures

4.2.2 Candidate Ranking Process

To identify whether the candidate has the same level of similarity to the examined user as the observed user, a ranking approach is used. In this approach, many similarity measures were combined into a single score for every candidate. The process involves the following:

1. **Assigning Weights:** Each of the similarity measures was given a weight depending on how significant it is. Due to their relevance in measuring

engagement, retweets and mentions were assigned higher priorities (weights each of 0.25).

2. **Data Normalization:** All scores were put on the same scale by using Min-Max Scaling to perform data normalization. As a result, the data could be properly compared between the different measures.
3. **Profile Similarity Adjustment:** Since lower Profile Similarity values indicate greater similarity, this measure was inverted to align with the other metrics.
4. **Calculating Composite Scores:** All the normalized scores were totaled, after which a common similarity score for each applicant was calculated.
5. **Ranking Candidates:** Candidates were ranked based on their composite scores, with lower scores indicating higher similarity to the examined user.

The final rankings, shown in Table 4.2, provide a comprehensive view of user similarity by combining interaction, content, and network-based metrics. This ranking process highlights the most significant candidates for further analysis or application.

Table 4.2. Candidates ranked based on composite score

Candidate	Composite Score	Rank
X1	1.327354	1
X10	0.974572	2
X2	0.952586	3
X4	0.924884	4
X6	0.630665	5
X8	0.608753	6
X9	0.471232	7
X3	0.434487	8
X5	0.408793	9
X7	0.133711	10
X11	0.051313	11

In conclusion, the combination of these tools allows us to view the similarities between users and determines which users are most similar to the one under study. They provide more insight into how users are connected, contributing to social network study and the design of personalized suggestions for users.

4.3 Evaluation

It is hard to determine if two users are alike, as similarity is based on personal opinions and ideas. On social media, how people act, what they post and how

they respond to others is what mainly shapes their evaluation. Since these platforms are always evolving and include many facets, it is hard to assess or measure the similarity between users.

4.3.1 Evaluation Approach

For our model to produce relevant and accurate results, we utilized human evaluation. While automated systems may give confusing results as found in previous studies, human evaluation is able to look at the different factors that make users similar. The approach required human review of the profiles, tweets and behavior of the examined user and each candidate user. After watching the candidates, human evaluators ranked the candidates according to how similar they were.

The rankings our model gave were compared with the rankings created by the human evaluators. The objective was to show which type of judgments best aligned with automated computations for observing the performance of similarity measures.

4.3.2 Evaluation Analysis

The evaluation revealed that the similarity elements in our model were efficient in representing how users are connected. Spearman's rank correlation coefficient is the statistic we use to evaluate how similar rankings are calculated by computers and by humans.

There was a strong and positive correlation shown by a Spearman's coefficient of 0.91 between the computer rankings and the human ratings. Since the two sets of rankings match well, it suggests our method and the metrics we use are effective. The following are the key observations:

- Accumulating different measures such as those based on interaction, content and network features, produced the highest-quality match results.
- While the results from the algorithm were nearly the same as the ones from humans, a few differences were noticed. Example: According to the model, Candidate X5 is ranked 9th, but people had him placed 5th. They show how difficult it is to come up with an algorithm that can account for all the influences that shape similarity assessments.

Table 4.3 shows the comparison between the computed rankings and the expert rankings given for each candidate user. As you can tell from the table, many of the candidates have consistent positions with only very minor fluctuations for some.

Table 4.3 Candidates computed rank versus human-evaluated rank

Candidate	Computed rank	Human evaluated rank
X1	1	1
X2	3	3
X3	8	9
X4	4	4
X5	9	5
X6	5	6
X7	10	10
X8	6	7
X9	7	8
X10	2	2
X11	11	11

4.3.3 Reflection on Subjectivity

Even though the study highlights a good relation between the two types of rankings, we must realize that such a match is based on personal opinions. Everyone's idea of similarity can be influenced by their own biases, knowledge and personal tastes. Since users act in different ways, it is important for automated systems to take in a wide variety of signals.

All in all, evaluating the model confirms that it accurately recognizes related users. Because the Spearman's coefficient is relatively high, the model performs well along with what humans judge, making it valuable for recommending products to people and analyzing social media interactions.

4.4 Discussion

This study investigated a number of means to assess user similarity on Twitter, sharing important information about the links between interactions, content and Twitter networks. By bringing together the different types of data, we hoped to find a good way to assess similarities between users and spot the main points where their features match or differ. The following are key findings:

It was found that Retweet Similarity and Interaction Count showed a very strong one-way relationship ($\rho=0.97$), underlining that users who interact a lot with a message tend to respond in many other ways as well. This finding emphasizes that interaction is an important factor for capturing relations between people and platforms.

A similar relationship ($\rho=0.87$) was identified between Hashtag Similarity and Common Friends, indicating that network connections are commonly reflected by

both people using the same hashtags. Still, although the metrics were created based on different ideas, their high agreement ($\rho \approx 0.82$) demonstrates that hashtag-based analysis accurately measures user resemblance.

Instead, Profile-Based Similarity did not perform as well in relation to comparisons with other features. Alternative factors might include what users do offline or like that just do not appear in their Twitter accounts. It looks like these results indicate that information from interactions and content work together, but signals from profiles often need further understanding to be included in a composite system of similarity.

Our evaluation of the model involved comparing the computed ranks to the human-made ranks and using Spearman's Rank Correlation Coefficient. There was a strong and consistent agreement between the two ranking methods ($\rho = 0.91$). Because the correlation is high, we know the composite similarity model is reliable and can accurately reflect how humans perceive user similarity.

While the findings are promising, there are several limitations to this study:

1. **Dataset Size and Diversity:** A database of only 11 candidate users was applied in the analysis for this case. With such a small number of participants, it becomes harder to generalize the research findings. By including more and different types of user actions, the model can be made more robust.
2. **Scalability Challenges:** Getting all the followers and followings of a user who has a huge social network is difficult when dealing with a lot of data. Moreover, needing to use less data from the API makes it more challenging to scale this system effectively.
3. **User-Specific Bias:** Relative to Studies: Using only one user could make calculating similarity for behaviors more influenced by his or her interaction signals and weights. It may not be effective when used in other kinds of situations.

While the study supports the approach taken, making changes to address the mentioned problems will help expand its use and practical use. They can improve applications for analyzing social networks, recommending users and identifying different online communities.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

In the evolving landscape of social media, measuring user similarity on platforms like Twitter has become a critical challenge with broad implications across sectors such as security, marketing, and social recommendations. Traditional methods relying on follower-following graphs are not sufficient to grasp the complexity of Twitter user behavior, forcing researchers to explore more complex techniques that integrate content, interactions, network structure, and user profiles.

The first part of the thesis surveyed the works that offer a wide variety of approaches, from graph-based models to machine-learning frameworks. In this respect, signal-based approaches have highlighted the advantage of integrating multiple signals from the platform, clustering, and similarity networks providing more insight into the grouping of users based on shared behavior, tastes, and sentiments. Each method presents specific benefits along with challenges such as computational complexity or the need for improved data labeling.

The second part of the study presented a comprehensive comparative assessment of three foundational models for identifying similarity between Twitter users: TSim, Characterizing and Detecting Similar Twitter Users, and Self-Similarity of Twitter Users. Using a range of signals, including those from social network structure, interactions and content, we checked how well each signal performed in identifying similar users. The purpose of the study was to group different types of indicators together to see if they could consistently provide accurate and useful results.

Data needed for the study was collected via the Twitter API which made it possible to find information about followers, following lists, posted tweets, interaction history, used hashtags and each user's profile data. The similarity score was calculated for each framework by adding up the scores from the signals.

The results point to a consistent agreement between the rankings made by humans and the model which had a Spearman Rank Correlation Coefficient of 0.91. According to the results, signals that people share or mention each other performed as the top predictors of a retweet. In addition to these, we used shared connections and the frequency of similar hashtags and topics to provide detailed context in our similarity analysis.

A key insight from the correlation analysis was the identification of redundancy and synergy among signal types. For example, the strong correlation between retweet similarity and interaction counts reinforced the hypothesis that active user engagement often coincides with broader interaction patterns. In contrast, when correlations are lower with profile-based views, it suggests that factors beyond Twitter impacts people's behavior. These findings affirm the hypothesis that integrating diverse signal types enhances the overall accuracy and robustness of user similarity estimation.

The framework presented positive results, but there are still ways to make it better and add to its functionality. The directions are aimed at making the framework

more appropriate and useful in practical, real-life situations.

1. **Adaptive Weighting Strategies:** In the present approach, each similarity is treated as per defined conditions, but their importance can shift in each scenario. Further investigations might involve adjusting the weights of signals depending on how important or useful they are for the individual user. The optimal weights used in similarity comparison can be trained using labeled similarity data in supervised machine learning models.
2. **Scalability and Large-Scale Testing:** For the framework to be useful, it must be checked with larger groups of people, including those with different ways of interacting and what they prefer to read. Tests with real systems like recommendation systems or user clustering can reveal the strength and effectiveness of the framework on a wide scale.
3. **Incorporating Temporal Dynamics:** This model does not consider how recently a user performed actions such as interactions, follows or shared content. In the future, temporal features could be integrated to follow the changes in user behavior, keeping the similarity rankings accurate.
4. **Multilingual and Regional Considerations:** Thinking about language and culture in the Twitter sample could also be an important step forward. If the framework included features for various languages and catered to different countries, it could analyze and compare user actions more easily in places with many languages.
5. **Applications in Social Media Personalization:** It is now important to turn the similarity rankings into useful solutions. Social intelligence tools might be applied to find influencers, target specific types of content, organize audiences into groups and make personal suggestions. For example, the similarity model could enable brands to choose suitable influencers and help social media sites provide more relevant content to users.
6. **Integration with Advanced Machine Learning Models:** Other studies might use deep learning or graph-based approaches to review the complex interactions and structure of users on the network. As a result, they might bring out features that traditional measures fail to detect, making the model more effective and varied.
7. **Evaluation Across Multiple Platforms:** The method used here could be used to assess Twitter user similarity and could also work for Instagram, LinkedIn or Facebook. By examining how users act on different social media sites we may find out how they manage their profiles and relationships online.

Bibliography

- [1] H. AlMahmoud and S. AlKhalifa, “TSim: a system for discovering similar users on Twitter,” *J Big Data*, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40537-018-0147-2.
- [2] A. Choumane and F. Yassin, “Characterizing and Detecting Similar Twitter Users,” in *2021 3rd IEEE Middle East and North Africa COMMUNICATIONS Conference, MENACOMM 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 25–30. doi: 10.1109/MENACOMM50742.2021.9678266.
- [3] M. Fatemi, K. Kucher, M. Laitinen, and P. Franti, “Self-Similarity of Twitter Users,” in *Proceedings of the 2021 Swedish Workshop on Data Science, SweDS 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/SweDS53855.2021.9638288.
- [4] M. A. U. Hasan, A. A. Bakar, and M. R. Yaakub, “Detecting Community Through User Similarity Analysis on Twitter,” in *Proceedings of the 2024 18th International Conference on Ubiquitous Information Management and Communication, IMCOM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IMCOM60618.2024.10418381.
- [5] A. Tundis, A. Jain, G. Bhatia, and M. Muhlhauser, “Similarity Analysis of Criminals on Social Networks: An Example on Twitter,” in *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, IEEE, Jul. 2019, pp. 1–9. doi: 10.1109/ICCCN.2019.8847028.
- [6] S. Tao, R. Qiu, Y. Ping, W. Xu, and H. Ma, “Making Explainable Friend Recommendations Based on Concept Similarity Measurements via a Knowledge Graph,” *IEEE Access*, vol. 8, pp. 146027–146038, 2020, doi: 10.1109/ACCESS.2020.3014670.
- [7] N. Shoeibi, N. Shoeibi, P. Chamoso, Z. Alizadehsani, and J. M. Corchado, “A Hybrid Model for the Measurement of the Similarity between Twitter Profiles,” *Sustainability (Switzerland)*, vol. 14, no. 9, May 2022, doi: 10.3390/su14094909.
- [8] S. Guo, K. Burghardt, V. Pantè, and K. Lerman, “SoMeR: Multi-View User Representation Learning for Social Media,” *arXiv preprint arXiv:2405.05275*, 2024.
- [9] ActiveGalaXy, “Tweets Targeting Isis,” Kaggle. Accessed: Dec. 20, 2024. [Online]. Available: <https://www.kaggle.com/datasets/activegalaxy/isis-related-tweets/data>
- [10] S. S. Sohail *et al.*, “Crawling Twitter data through API: A technical/legal perspective,” *arXiv preprint arXiv:2105.10724*, 2021.
- [11] L. Luceri, V. Pantè, K. Burghardt, and E. Ferrara, “Unmasking the Web of Deceit: Uncovering Coordinated Activity to Expose Information Operations on Twitter,” in *Proceedings of the ACM Web Conference 2024*, New York, NY, USA: ACM, May 2024, pp. 2530–2541. doi: 10.1145/3589334.3645529.
- [12] A. C. Nwala, A. Flammini, and F. Menczer, “A language framework for modeling social media account behavior,” *EPJ Data Sci*, vol. 12, no. 1, Dec. 2023, doi: 10.1140/epjds/s13688-023-00410-9.
- [13] A. Anwala, “General language behavior: Control dataset,” GitHub. Accessed:

- Dec. 20, 2024. [Online]. Available: <https://github.com/anwala/general-language-behavior/tree/main/coordination-detect/control-dataset>
- [14] J. Zhang, J. Fei, X. Song, and J. Feng, “An Improved Louvain Algorithm for Community Detection,” *Math Probl Eng*, vol. 2021, pp. 1–14, Nov. 2021, doi: 10.1155/2021/1485592.
 - [15] L. Ge and S. Chen, “Exact Dynamic Time Warping calculation for weak sparse time series,” *Applied Soft Computing Journal*, vol. 96, Nov. 2020, doi: 10.1016/j.asoc.2020.106631.
 - [16] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, “WTF,” in *Proceedings of the 22nd international conference on World Wide Web*, New York, NY, USA: ACM, May 2013, pp. 505–514. doi: 10.1145/2488388.2488433.
 - [17] D. Downey and B. Malin, “A Semantic Cover Approach for Topic Modeling,” 2019. doi: 10.18653/v1/S19-1011.



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis _____

Total Pages _____ Name of the Scholar _____

Supervisor (s)

(1) _____

(2) _____

(3) _____

Department _____

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: _____ Similarity Index: _____, Total Word Count: _____

Date: _____

Candidate's Signature

Signature of Supervisor(s)

Thesis-twitter-02.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:96536638

Submission Date

May 18, 2025, 9:57 PM GMT+5:30

Download Date

May 18, 2025, 10:00 PM GMT+5:30

File Name

Thesis-twitter-02.pdf

File Size

631.3 KB

38 Pages

10,511 Words

66,215 Characters





8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups


-  **47 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 6%  Internet sources
- 2%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

-  **Replaced Characters**
38 suspect characters on 8 pages
Letters are swapped with similar characters from another alphabet.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 47 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 6% Internet sources
- 2% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.dtu.ac.in:8080	1%
2	Internet	de.overleaf.com	<1%
3	Submitted works	dtusimilarity on 2024-05-29	<1%
4	Internet	arxiv.org	<1%
5	Submitted works	Coventry University on 2025-04-07	<1%
6	Publication	Md Ahsan Ul Hasan, Azuraliza Abu Bakar, Mohd Ridzwan Yaakub. "Detecting Com...	<1%
7	Internet	www.slideshare.net	<1%
8	Submitted works	University of Wollongong on 2023-12-08	<1%
9	Submitted works	Lebanese University on 2021-03-27	<1%
10	Internet	link.springer.com	<1%

11	Internet	thesai.org	<1%
12	Submitted works	University of Westminster on 2025-04-07	<1%
13	Internet	vdocuments.mx	<1%
14	Submitted works	University of Northampton on 2025-05-13	<1%
15	Submitted works	University of Hertfordshire on 2024-09-02	<1%
16	Internet	educationdocbox.com	<1%
17	Publication	Bohrer, Forest I. "Gas sensing mechanisms in chemiresistive metal phthalocyanin..."	<1%
18	Submitted works	Indian Institute of Technology Roorkee on 2022-05-10	<1%
19	Submitted works	University of Sheffield on 2012-05-02	<1%
20	Internet	d-nb.info	<1%
21	Internet	www.researchgate.net	<1%
22	Publication	Ali Choumane, Fatima Yassin. "Characterizing and Detecting Similar Twitter User..."	<1%
23	Publication	Shaohua Tao, Runhe Qiu, Yuan Ping, Woping Xu, Hui Ma. "Making Explainable Fri..."	<1%
24	Internet	etd.lib.metu.edu.tr	<1%

25	Publication	"Proceedings of Second Doctoral Symposium on Computational Intelligence", Spr...	<1%
26	Publication	"The AI Cleanse: Transforming Wastewater Treatment Through Artificial Intellige...	<1%
27	Publication	Fatima Dakalbab, Manar Abu Talib, Omnia Abu Waraga, Ali Bou Nassif, Sohail Abb...	<1%
28	Publication	Masoud Fatemi, Kostiantyn Kucher, Mikko Laitinen, Pasi Franti. "Self-Similarity of ...	<1%
29	Submitted works	University of Birmingham on 2016-01-17	<1%
30	Submitted works	University of Wales Institute, Cardiff on 2016-04-19	<1%
31	Internet	anyflip.com	<1%
32	Internet	digital.library.adelaide.edu.au	<1%
33	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
34	Submitted works	paper	<1%
35	Internet	www.mdpi.com	<1%

Thesis-twitter-02.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:96536638

Submission Date

May 18, 2025, 9:57 PM GMT+5:30

Download Date

May 18, 2025, 10:00 PM GMT+5:30

File Name

Thesis-twitter-02.pdf

File Size

631.3 KB

38 Pages

10,511 Words

66,215 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.



APPENDIX A

A.1 LIST OF PUBLICATION

1. Vidushi Jain, Priyanka Arora & Dr. Sonika Dahiya (2025). A Comprehensive Review of Methods for Measuring User Similarity on Twitter, 8th International Conference on Innovative Computing and Communication (ICICC-2025). **[Scopus Indexed] [Accepted]**
2. Vidushi Jain (2025) & Dr. Sonika Dahiya. Detecting Similar Twitter Users : A Multi-Signal Comparative Analysis, 6th International Conference on Data Analytics & Management (ICDAM-2025). **[Scopus Indexed] [Accepted]**

A.2 PAPER ACCEPTANCE PROOF

A.2.1 1st Conference Acceptance Proof:



Vidushi Jain <vishi13jain@gmail.com>

ICICC 2025: Paper Notification for Paper ID 312

4 messages

ICICC 2025 <icicc.ui@gmail.com>
To: Vishi13jain <vishi13jain@gmail.com>

Fri, Nov 1, 2024 at 2:00 PM

8th International Conference on Innovative Computing and Communication (ICICC-2025) - A Flagship Conference

Dear Author(s),

Greetings from - ICICC 2025!

We congratulate you that your paper with submission ID '312' and Paper Title '**A Comprehensive Review of Methods for Measuring User Similarity on Twitter**' has been accepted for publication in the **Proceedings of ICICC 2025- Springer LNNS Series** [Indexing: zbMATH, Scopus and Web of Science - Proposed]. This acceptance means your paper is among the top 20% of the papers received/reviewed.

To secure your spot at this highly anticipated event and to opt for early bird registration, we urge you to complete your registration as soon as possible (<https://icicc-conf.com/registrations>).

You are requested to do the registration as soon as possible and submit the following documents to icicc.ui@gmail.com at the earliest.

1. Final Camera-Ready Copy (CRC) as per the springer format- (See <https://icicc-conf.com/downloads>)
2. Copy of e-receipt of registration fees. (For Registration, see <https://icicc-conf.com/registrations>).
3. The final revised copy of your paper should also be uploaded via Microsoft CMT.

The reviewers' comments are given at the bottom of this letter, please improve your paper as per the reviewers comments. **While preparing the final CRC manuscript, kindly check the following link under the download section tab named manuscript guidelines:**

<https://icicc-conf.com/registrations>

and it is suggested to cite the relevant latest papers matching the area of your current research paper.

The paper prior to submission should be checked for plagiarism from licensed plagiarism softwares like Turnitin/iAuthenticate etc. The similarity content should not exceed 15%.

Pay registration fees via online portal:

A.2.2 2nd Conference Acceptance Proof:



Vidushi Jain <vishi13jain@gmail.com>

ICDAM 2025: Paper Notification for Paper ID 1227

2 messages

ICDAM 2025 <icdam.conf@gmail.com>
To: Vishi13jain <vishi13jain@gmail.com>

Tue, May 6, 2025 at 4:55 PM

6th International Conference on Data Analytics & Management (ICDAM-2025)!

Dear Author(s),

Greetings from **ICDAM 2025**!

We congratulate you that your paper with submission ID **1227** and Paper Title '**Detecting Similar Twitter Users : A Multi-Signal Comparative Analysis**' has been accepted for publication in the Springer LNNS series [Indexing: SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago; All books published in the series are submitted for consideration in Web of Science]. This acceptance means your paper is among the top 20% of the papers received/reviewed. We urge you to complete your registration immediately to secure your spot at this highly anticipated event. **Please submit the revised paper by May 15th 2025, and complete the registration by May 15th, 2025 , to receive the early bird offer. We are left with very few slots. On the website it's 30th April 2025 but you have got the late acceptance , so the deadline for the registration is 15th May 2025.**

Please register as soon as possible and submit the following documents to icdam.conf@gmail.com as soon as possible.

1. Final Camera-Ready Copy (CRC) as per the springer format. (See <https://icdam-conf.com/downloads>)
2. Copy of e-receipt of registration fees. (For Registration, see <https://icdam-conf.com/registrations>)
3. The final revised copy of your paper should also be uploaded via Microsoft CMT.

Note : Standard Paper size – 10-12 pages. Over length of more than 12 pages, paper charges USD 20 per extra page

The reviewers comments are given at the bottom of this letter, please improve your paper as per the reviewers comments. While preparing the final CRC manuscript, kindly check the following google link of proceedings of the previous International Conference on Data Analytics and Management:

<https://link.springer.com/conference/icdam>

and it is suggested to cite the relevant latest papers matching the area of your current research paper.

A.3 INDEXING OF CONFERENCE PROOF

A.3.1 1st Conference

5/19/25, 2:31 PM

ICICC | International Conference on Innovative Computing and Communication


ICICC

 INTERNATIONAL CONFERENCE ON INNOVATIVE
COMPUTING AND COMMUNICATION

**9th
INTERNATIONAL
CONFERENCE ON
INNOVATIVE
COMPUTING AND
COMMUNICATION
(ICICC-2026)**



ORGANISED BY:
SHAHEED SUKHDEV
COLLEGE OF BUSINESS
STUDIES, UNIVERSITY OF
DELHI, NEW DELHI
IN ASSOCIATION WITH
NATIONAL INSTITUTE OF
TECHNOLOGY PATNA &
UNIVERSITY OF
VALLADOLID SPAIN
6th-7th FEBRUARY
2026

ICICC 2025

The eighth version of the International Conference in Innovative Computing and Communication (ICICC-2025) was organized at Shaheed Sukhdev College of Business Studies in association with the National Institute of Technology Patna and the University of Valladolid Spain, on 14-15 February 2025 at New Delhi, India. ICICC-2025 received 2000 papers from approximately 6000 plus authors and a total of 400 papers were accepted with an acceptance ratio of 20%. All accepted papers were published in Springer's Lecture Notes on Networks and Systems, a Scopus-indexed series. A total of 750 participants attended the conference including authors, keynotes, delegates, academicians, and industry experts. ICICC-2025 received papers from 35 countries. ICICC-2025 was organized in hybrid mode.

Important Dates

<https://icicc-conf.com/icicc25>

1/4

A.3.2 2nd Conference

5/19/25, 2:30 PM

ICDAM | International Conference on Data Analytics and Management



6th International Conference on Data Analytics & Management (ICDAM-2025)
ICDAM-2025 Theme: Data Analytics with Computer Networks
Organized By: London Metropolitan University, London, UK (Venue Partner)
in association with
WSG University, Bydgoszcz Poland, Europe
&
Portalegre Polytechnic University, Portugal, Europe
&
SGW Management Institute
Date: 13th - 15th June, 2025
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)

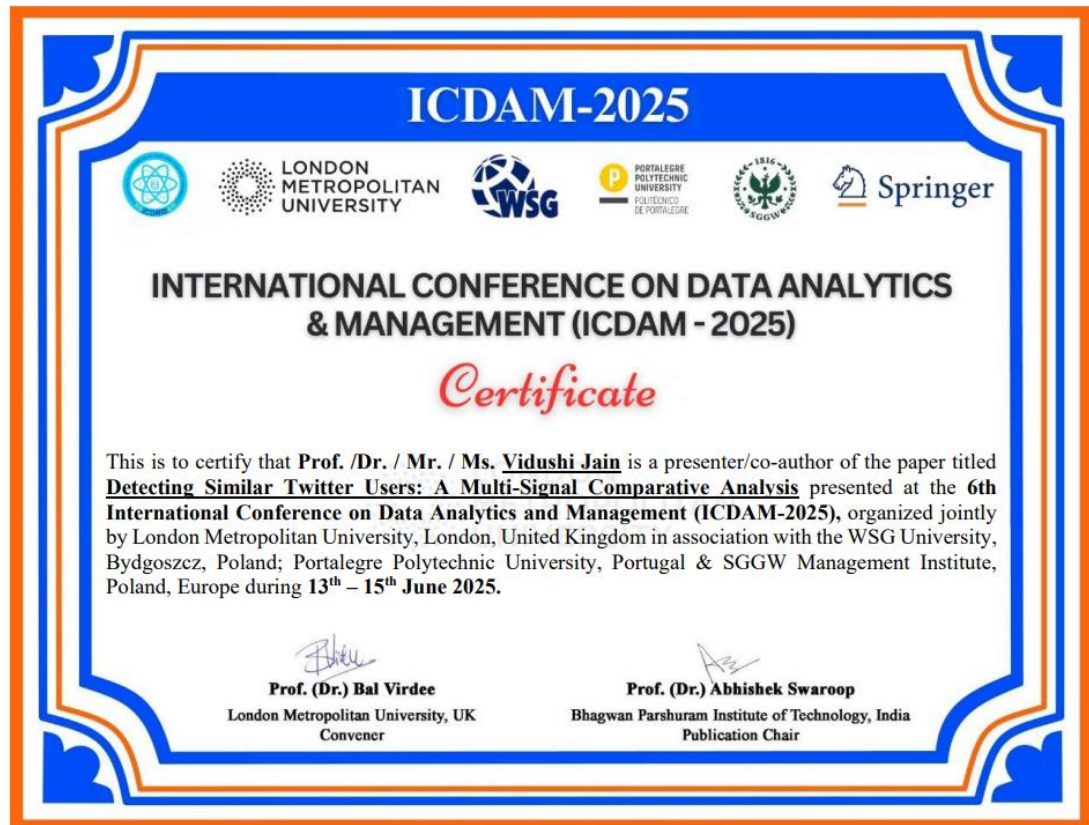


A.4 CONFERENCE CERTIFICATE

A.4.1 1st Conference Certificate



A.4.2 2nd Conference Certificate



A.5 CONFERENCE PAPER REGISTRATION PROOF

A.5.1 1st Conference Registration Proof



Vidushi Jain <vishi13jain@gmail.com>

Subject: Final Submission (Springer -Paper ID 312) Instructions for Accepted Papers – ICICC 2025

1 message

ICICC 2025 <icicc.ui@gmail.com>
To: Vishi13jain <vishi13jain@gmail.com>

Sun, Dec 15, 2024 at 1:25 PM

Dear Authors,

Congratulations once again on the acceptance and registration of your paper -Paper ID 312 and Paper Title A Comprehensive Review of Methods for Measuring User Similarity on Twitter for ICICC 2025. We are delighted to have you as part of this esteemed conference.

To proceed, we request you to upload the **final version** of your paper in both **Word** and **PDF** formats (Google Form Link : <https://docs.google.com/forms/d/e/1FAIpQLSe0NNTr12nLqwTA9O-msxzBn5p1oNhCpoXcJo7iuxHW3WSyGQ/viewform?usp=header>) , incorporating all the reviewer suggestions. Please ensure that the final version adheres to the following guidelines:

1. Paper Length:

- The paper must be a minimum of **8 pages** and a maximum of **12 to 15 pages**.

2. Authors and Affiliations:

- Include the names and affiliations of all authors.
- Mark the corresponding author with an asterisk (*).

3. Tables:

- Ensure all tables are in an **editable format**.
- Provide appropriate captions and number all tables sequentially.

4. Images:

- Include high-quality images.
- Ensure all images are properly cited within the paper.

5. Captions and Numbering:

- Provide clear captions for all figures and tables.
- Number all figures and tables sequentially.

6. Self-Citations:

- Limit self-citations to a maximum of **three references**.

Important Note:

- Do not color or highlight changes in the final version of your paper.

We appreciate your adherence to these guidelines to ensure a smooth publication process. Should you have any queries or require assistance, feel free to reach out to us.

Looking forward to your submission.

Warm regards,

Program Chair,

ICICC 2025

A.5.2 2nd Conference Registration Proof



Vidushi Jain <vishi13jain@gmail.com>

Submission of Final CRC and Registration Receipt for Paper ID 1227 in ICDAM 2025

ICDAM Conf <icdam.conf@gmail.com>
To: Vidushi Jain <vishi13jain@gmail.com>

Thu, May 15, 2025 at 8:26 PM

Dear Author(s),

Greetings from ICDAM 2025 Organizing Committee!

We are pleased to inform you about the final presentation arrangements for your accepted & registered paper (Paper ID: 1227). Please find attached the **Presentation Template** and **Zoom background poster** (for *online* presenters only).

Important Instructions:

- **Presentation Time:** 08 minutes for the presentation, followed by 02 minutes for the Q&A session.
- **Mode:** Details for both *Online* and *Offline* presentations are covered.
- **Presentation Schedule:** The detailed schedule and meeting links will be shared soon. Kindly **do not email requesting changes** to the scheduled day.
- **Template Usage:** All presenters are requested to prepare their presentation strictly as per the attached template.
- **Zoom Background (Online Only):** Use the attached virtual background during your online presentation. (*Offline presenters may ignore this.*)

We appreciate your cooperation and look forward to your participation in ICDAM 2025.

Best Regards,
Conference Convener
ICDAM 2025 Organizing Committee

[Quoted text hidden]

2 attachments



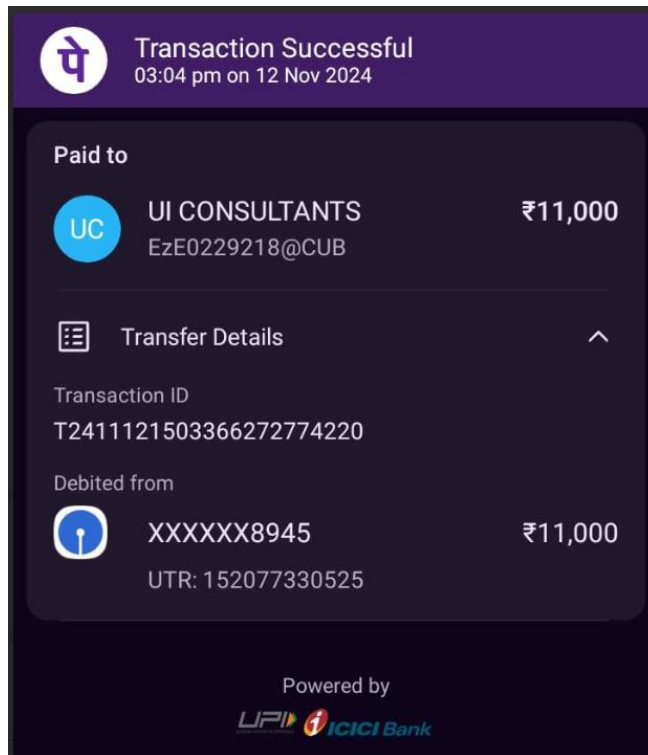
BACKGROUND POSTER ICDAM 2025.jpg
193K



ICDAM_2025_PPT_Template-SPRINGER.pptx
472K

A.6 CONFERENCE PAPER REGISTRATION RECEIPT

A.6.1 1st Conference Receipt



A.6.2 2nd Conference Receipt

UNIVERSAL INOVATORS



Payment Receipt

Transaction Reference: pay_QVDtBQUHCa4VSO

This is a payment receipt for your transaction on ICDAM 2025

AMOUNT PAID ₹ 15,250.00

ISSUED TO
vishi13jain@gmail.com
+917906654105

PAID ON
15 May 2025

DESCRIPTION	UNIT PRICE	QTY	AMOUNT
Service Charges	₹ 250.00	1	₹ 250.00
Registration Fee	₹ 15,000.00	1	₹ 15,000.00
Total			₹ 15,250.00
Amount Paid			₹ 15,250.00

DECLARATION

We/I hereby certify that the work which is presented in the Major Project-II/Research Work entitled _____ in fulfilment of the requirement for the award of the Degree of Bachelor/Master of Technology in _____ and submitted to the Department of _____, Delhi Technological University, Delhi is an authentic record of my/our own, carried out during a period from _____, under the supervision of _____.

The matter presented in this report/thesis has not been submitted by us/me for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/ SCI expanded/SSCI/Scopus indexed journal OR peer reviewed Scopus indexed conference with the following details:

Title of the Paper:

Author names (in sequence as per research paper):

Name of Conference/Journal:

Conference Dates with venue (if applicable):

Have you registered for the conference (Yes/No)?:

Status of paper (Accepted/Published/Communicated):

Date of paper communication:

Date of paper acceptance:

Date of paper publication:

Student(s) Roll No., Name and Signature

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the students is correct.

Place: _____

Supervisor Name and Signature

Date: _____

**NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/
PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF
(Conference Website OR Science Direct in case of Journal Publication).**