

Multi-View User Similarity Analysis on Twitter

A THESIS
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF
MASTERS OF TECHNOLOGY
IN
DATA SCIENCE

Submitted by
MUSKAN JAIN
(2K24/DSC/02)

Under the supervision of
Dr. SONIKA DAHIYA
(ASSISTANT PROFESSOR)



DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

MAY, 2026

ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to **Dr. Sonika Dahiya**, Assistant Professor, Department of Software Engineering, Delhi Technological University, for her continuous guidance, support, and mentorship throughout the preparation of this thesis. Her insights and expertise were instrumental in shaping the direction of my research and explaining all the tasks to be accomplished, emphasising the importance of this work as well as its relevance to real-world applications.

I would also like to thank all the teachers and students of the Department of Software Engineering, Delhi Technological University, for providing the necessary resources, continuous support, and valuable suggestions that facilitated the completion of this research. Lastly, I would like to thank all my family and friends for providing continuous support and motivation during the process of this study.

MUSKAN JAIN

2K24/DSC/02

Department of Software Engineering
Delhi Technological University



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I **MUSKAN JAIN, Roll No. 2K24/DSC/02** hereby certify that the work which is being presented in the thesis entitled "*Multi-View User Similarity Analysis on Twitter*" in partial fulfillment of the requirements for the award of the 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 January 2026 to May 2026 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



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **MUSKAN JAIN** (enrollment no. 2K24/DSC/02) has carried out their search work presented in this thesis entitled “**Multi-View User Similarity Analysis on Twitter**” for the award of **Master of Technology** from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature

Dr. Sonika Dahiya

Assistant Professor

Department of Software Engineering

Delhi Technological University

Date: _____

Multi-View User Similarity Analysis on Twitter

Muskan Jain

ABSTRACT

With the rise of social media sites like Twitter, there is a huge amount of user-generated data available, making possible various applications like recommendations, community finding, user profiling, and targeted marketing. User similarity on Twitter can be a difficult problem to solve because social media data consists of a variety of components, including text data, behaviors, user profiles, and connections. The traditional method of solving the user similarity problem by using only one type of data does not work effectively anymore.

This thesis offers a systematic study and comparative analysis of user similarity techniques on Twitter. The review covers previous research in this area, classifies techniques based on the type of their work (content-based, graph-based, signal-based, or embedding-based), and discusses main difficulties of this research such as data sparsity, noise in text, scalability, and lack of benchmarks. It also notes the emerging trend to move away from feature-based techniques towards hybrid and representation learning-based techniques.

The experimental comparison study compares three types of similarity models: those based on features, representation learning, and metadata. The feature-based method is based on semantic, behavioral, and structural similarity metrics, while the representation learning method relies on the use of Sentence-BERT embeddings and temporal behavioral modeling to determine latent relationships between users. Metadata-based similarity approach depends on profile information, textual similarity, and network-based methods.

From the experiments conducted, it can be seen that the similarities between users in social media sites are multi-dimensional in nature and therefore cannot be captured by just one method alone. The feature-based approach is more interpretable, representation learning offers better latent connections, and the metadata-driven approach performs better in terms of classification. In conclusion, it can be said that hybrid and multi-view approaches for measuring similarity are more effective.

TABLE OF CONTENTS

Title	Page No.
Acknowledgement	ii
Candidate's Declaration	iii
Certificate	iv
Abstract	v
Contents	vi
List of Tables	viii
List of Figures	ix
List of Abbreviations	x
CHAPTER 1: INTRODUCTION	1
1.1 Problem Statement	1
1.2 Research Objectives	2
1.3 Overview of the Study	2
CHAPTER 2: REVIEW OF USER SIMILARITY APPROACHES ON TWITTER	4
2.1 Introduction	4
2.2 Search Methodology	5
2.2.1 Search Process	5
2.2.2 Selection Criteria	6
2.3 User Similarity Approaches on Twitter	6
2.3.1 Content-Based Approaches	6
2.3.2 Graph-Based Approaches	7
2.3.3 Signal-Based Approaches	7
2.3.4 Embedding-Based Approaches	8
2.4 Datasets and Features Used	9
2.5 Evaluation Metrics	10
2.6 Challenges, Research Trends, and Research Gaps	11
2.7 Summary	12
CHAPTER 3: COMPARATIVE ANALYSIS OF USER SIMILARITY APPROACHES ON TWITTER	14
3.1 Dataset Collection	14
3.2 Similarity Approaches	15
3.3 Performance Metrics	15
3.4 Results and Comparative Analysis	16
3.4.1 Feature-Based Similarity Results	16
3.4.2 Representation Learning Results	16
3.4.3 Metadata-Driven Similarity Results	17
3.4.4 Comparative Analysis of All Approaches	17
3.5 Discussion	18
3.6 Summary	18

CHAPTER 4: CONCLUSION, LIMITATIONS AND FUTURE SCOPE	20
4.1 Conclusion	20
4.2 Limitations	20
4.3 Future Scope	21
Bibliography	22

LIST OF TABLES

2.1	Comparative Summary of User Similarity Approaches on Twitter . . .	9
2.2	Datasets Used in Selected Twitter User Similarity Studies	10
2.3	Evaluation Metrics and Performance Summary of Selected Studies . .	11
3.1	Comparative Summary of Implemented Similarity Approaches	15
3.2	Comparison of Different Similarity Approaches	17

LIST OF FIGURES

1.1	Overview of the Two-Phase Research Study	3
2.1	Workflow of the systematic literature review for user similarity methods on Twitter.	5

LIST OF ABBREVIATIONS

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DTW	Dynamic Time Warping
F1-Score	Harmonic Mean of Precision and Recall
GNN	Graph Neural Network
GPU	Graphics Processing Unit
JSON	JavaScript Object Notation
k-NN	k-Nearest Neighbors
LDA	Latent Dirichlet Allocation
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
ROC-AUC	Receiver Operating Characteristic – Area Under Curve
SBERT	Sentence Bidirectional Encoder Representations from Transformers
SLR	Systematic Literature Review
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TSim	Twitter Similarity Model
URL	Uniform Resource Locator
LSTM	Long Short-Term Memory

Chapter 1

INTRODUCTION

Millions of people create large amounts of text data, behavioral data, and interaction data through tweeting, retweeting, replying, mentioning, hashtagging, and social relationships on a daily basis. Such large scale user-generated data can provide many interesting possibilities for analyzing social networks and creating intelligent data-driven systems.

User similarity involves finding users who are alike in terms of their interests, behaviors, interactions, or social network features. Applications of user similarity include recommendation systems, community detection, targeted marketing, user profiling, and security monitoring [1, 2]. However, measuring the similarity of Twitter users is a difficult problem due to the heterogeneity and dynamism of user information. User similarity cannot be computed based on text alone because users are different from each other with respect to their behavioral aspects, temporal aspects, profile characteristics, and interactions [3, 4].

1.1 Problem Statement

The problem of user similarity analysis on Twitter is still an open issue due to the heterogeneous and dynamic nature of the social media data. There exist various types of approaches to measuring user similarity on social media platforms, which rely on textual data, interactions, behaviors, profile properties, and network structure. However, most prior works concentrate on just one category of similarity measures and do not address the multidimensional nature of the user behavior on social media [5, 3].

The content-based approaches are effective for detecting semantic similarities between users based on tweets, hashtags, and profile descriptions. However, they are very sensitive to noise and sparsity of the textual data [5]. The graph-based and interaction-based approaches detect the network relationships between users, but ignore the semantics and context of the tweets [2]. On the other hand, the signal-based approaches are heavily dependent on manually defined features and weighting schemes

[1]. Finally, the recent representation learning approaches offer improved ability to learn latent relationships in the data, but they require large-scale data sets and are computationally expensive [6, 7].

In addition to these limitations, another important drawback of the existing literature on user similarity analysis is the absence of a systematic comparison of different types of user similarity approaches under identical experimental conditions.

1.2 Research Objectives

1. To study and classify user similarity techniques into content-based, graph-based, signal-based, and embedding-based techniques depending on their approach and characteristics.
2. To implement and compare feature-based, representation learning-based, and metadata-based techniques by applying them to Twitter data collected through API mining and publicly available datasets.
3. To examine semantic, behavioral, profile-based, and network-based similarities in finding similar users on Twitter.
4. To test the performance of various similarity techniques through classification and similarity metrics.
5. To highlight the advantages, limitations, and practical implications of various user similarity techniques for future research.

1.3 Overview of the Study

This thesis presents a comparative analysis of user similarity techniques in the context of the social media network Twitter by using two parallel approaches: a systematic review of existing literature and comparative experiments. The proposed methodology is described in Figure 1.1.

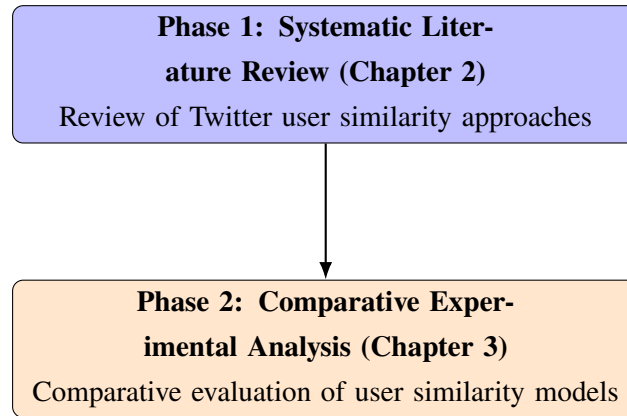


Figure 1.1: Overview of the Two-Phase Research Study

Phase 1: Systematic Review (Chapter 2): The first phase involves conducting a systematic literature survey of user similarity approaches currently available for Twitter. Research papers in this area are reviewed and classified based on four major categories of user similarity algorithms, namely, content-based, graph-based, signal-based, and embedding-based methods. This review considers commonly utilized data sources, features, evaluation criteria, benefits, drawbacks, and challenges related to various user similarity measures.

Phase 2: Comparative Study (Chapter 3): In the second phase, a comparative experimental study is conducted for three types of similarity approaches for users, namely: feature similarity approach, representation learning similarity approach, and metadata similarity approach. For the experiment, Twitter data is extracted from the API and public data sources. Different types of similarities based on semantics, behavior, profile, and network are studied by comparing them based on classification and similarity measurement techniques.

Chapter 2

REVIEW OF USER SIMILARITY APPROACHES ON TWITTER

This chapter provides a systematic literature review of various methods that have been used to measure similarity among Twitter users. Various research studies conducted between the years 2015 and 2025 are critically reviewed and focused mainly on content based, graph based, signal based, and embedding based similarity measurement methods.

2.1 Introduction

Many applications have been developed from such big data including recommendation systems, community detection, advertising, and user profiling [1, 8]. The identification of user similarities is one of the critical aspects of social network analysis that depends on users' interests, behaviors, network structure, and interactions.

There might be various ways to estimate similarity depending on factors like tweets, hashtags, activity time, profile features, followers, mentions, retweets, and behavioral patterns [3, 4]. Prior research was mostly limited to content-based approaches based on textual elements like tweets and hashtags. Later work incorporated graph-based similarity measures using social network structure and interactions. Recently proposed approaches involve signal-based and embedding-based methods that allow for incorporating both semantic, behavioral, and structural characteristics through representation learning [6, 7].

There are still a number of challenges related to measuring user similarity. Most existing methods are not holistic enough and cover only some specific aspects of user similarity. Moreover, textual noise, sparse interactions, scalability problems, evolving user behaviors, and a lack of benchmark datasets complicate the comparison of different approaches [9]. Therefore, there is a tendency toward developing hybrid, multi-view, and representation learning-based approaches.

This chapter provides an overview of existing Twitter user similarity approaches and categorizes them into four types: content-based, graph-based, signal-based, and embedding-based. Besides, the chapter explores datasets, feature extraction techniques, evaluation metrics, research challenges, and future research directions associated with user similarity on Twitter.

2.2 Search Methodology

The SLR approach is used in this research paper to examine the various methods employed to measure the similarities between users on Twitter. Literature search, study selection, full-text analysis, data extraction, and reporting are some of the steps involved in conducting a systematic literature review. The selected studies were then examined depending on their similarity measurement methods, datasets, features, evaluation criteria, advantages, limitations, and research contributions. The overall workflow of the systematic literature review process is illustrated in Figure 2.1.

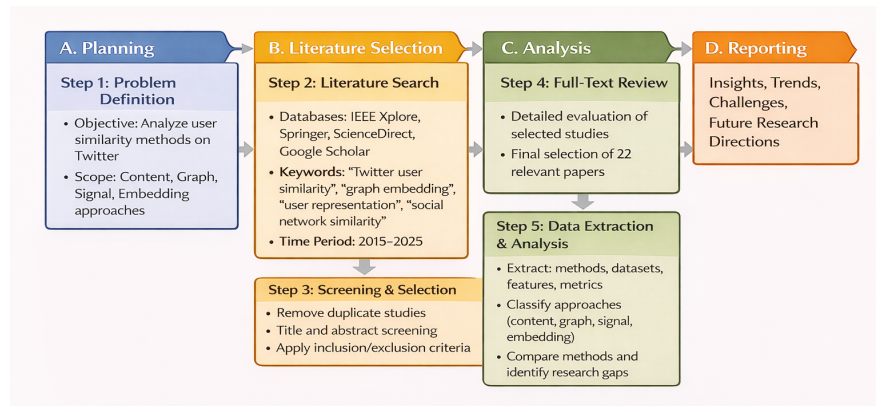


Figure 2.1: Workflow of the systematic literature review for user similarity methods on Twitter.

2.2.1 Search Process

A literature search was carried out for techniques to find similarities between Twitter users using key academic databases such as IEEE Xplore, Springer, ScienceDirect, ACM Digital Library, and Google Scholar. Different keywords were employed for each of the searches, which included Twitter user similarity, graph embedding, social network similarity, and user representation learning. The studies covered in this review were those published between 2015 and 2025.

2.2.2 Selection Criteria

In the first stage, titles and abstracts of the gathered papers were reviewed to find out those which were about the similarity analysis of users in Twitter. In the second stage, the complete texts of the identified papers were analyzed in order to assess their methodologies, data sets, similarity techniques, and contributions to the field.

Papers addressing similarity analysis using content-based, graph-based, signal-based, and embedding-based techniques were considered. Duplicate papers, irrelevant papers, non-English papers, and papers without detailed information on methodology were excluded from consideration.

2.3 User Similarity Approaches on Twitter

The existing techniques for estimating user similarities in Twitter can be classified into four main categories, content-based, graph-based, signal-based, embedding-based techniques. Key difference between these techniques is the type of information employed to estimate the similarity measure. While earlier research focused on the use of textual and profile information, more recent research employs a combination of several types of information to better estimate the similarity measure [3, 8].

Each technique has its strengths and weaknesses, which will be discussed in further subsections.

2.3.1 Content-Based Approaches

The content-based approach involves estimating the similarity between Twitter users based on the text data generated by tweets, hashtags, profile descriptions, and other forms of user-generated content. The underlying logic of content-based approaches is that users who share the same themes and textual patterns are expected to have common interests and behavior patterns. Initially, content-based studies employed traditional text similarity measures like TF-IDF and cosine similarity to measure similarities between tweets and user profiles [1, 8].

In order to enhance semantic understanding, some studies used ontology-based and semantic similarity measures that examine contextual relations between words, hashtags, mentions, and URLs [5]. Additionally, some methods integrated textual data with machine learning models for user profiling and classification to identify users with common interests and behaviors [4]. Hypernym-based feature generation and tweet similarity calculation have also been shown to enhance the process of estimating

semantic similarity [10]. Content-based approaches are typically easy to implement, understandable, and fast, which makes them appropriate for large datasets.

There are several disadvantages associated with the use of content-based methods for similarity estimation. First, tweets are rather short, noisy, informal, and highly dynamic, thus limiting their ability to capture user similarities. Second, the content-based approach does not provide any information about social ties and interactions among users since it relies mostly on textual data. Therefore, content-based approaches are commonly integrated with interaction/network-based features.

2.3.2 Graph-Based Approaches

The graph-based approaches use the Twitter network in terms of a social graph where the nodes are the users and the edges are interactions such as followers, followings, mentions, replies, and retweets. The similarity is estimated based on structural relations and interaction patterns among users. As opposed to the content-based approaches, the graph-based approaches do not depend only on the textual information but take into account social interactions and community structures [2, 3]. Graph neural network and node embedding models have been used to perform community detection and structural analysis in large-scale social graphs [11].

A number of research works rely on various graph measures including common neighbors, shortest path distance, interaction frequency, and community detection to measure user similarity [4]. Moreover, some recently proposed approaches consider graph representation learning to learn structural relations from Twitter social graphs [7]. Graph embedding techniques have also been used to encode Twitter interaction graphs for user classification and political closeness assessment [12].

The strength of the graph-based approaches lies in their capability of capturing social relations and interaction patterns that cannot be captured using textual analysis. On the other hand, the graph-based approaches ignore the semantic information in the tweets and profiles. Furthermore, large-scale social graphs pose computational complexities due to graph building and similarity estimation tasks.

2.3.3 Signal-Based Approaches

The main idea behind signal-based approaches is that these techniques estimate similarity between users in Twitter based on signals gathered through behavioral patterns and interactions among Twitter users. The most common signals used in signal-based methods include followers, following, retweeting, mentioning, replying, hashtag, likes, number of interactions between users, and profile information. Unlike the

previous approach, signal-based approaches use several behavioral factors to estimate the similarity of users in a flexible and effective way [1, 8]. Several researchers apply weighted combination techniques using different types of signals to define relations among Twitter users. For instance, the use of TSim methods allows combining several interaction features, assigning different weights to each factor to calculate similarity score between two users [1].

The main advantage of signal-based approaches is that these techniques allow combining different sources of information while being capable of controlling weights assigned to each feature in similarity computation process. Signal-based approaches are usually much more effective compared to single-sourced approaches since they consider both behavioral and interactional aspects of user similarity assessment. Nevertheless, effectiveness of these techniques is heavily dependent on correct feature selection and their weights assignment.

2.3.4 Embedding-Based Approaches

Approaches based on embeddings are some of the most recent techniques used to assess similarity among Twitter users. They leverage the learned low-dimensional representations of user behavior to capture similarity. In these approaches, the vector representations of users are learned by considering various types of data from social media. The similarities among users are estimated by measuring the distance between the vector representations of the users. Cosine similarity is widely used in embedding-based approaches for user similarity assessment [6, 7]. Social embedding models such as SocialVec are used to capture low-dimensional embeddings from the social context of co-following relationships on Twitter [13].

Some of the embedding-based approaches include the use of graph embedding and representation learning approaches like Node2Vec, graph neural networks, and multi-view representation learning. Other more sophisticated embedding-based approaches consider temporal activities, semantic information, and interaction patterns to capture user similarities within the dynamic social networks [12, 14, 15]. Heterogeneous Information Network embedding for modeling interactions and recommendations on multi-relational Twitter data sets is another area where recent work using embedding techniques has been carried out [16]. Further research has explored contrastive learning approaches for creating robust node embeddings using different forms of graph augmentation [14]. In general, embedding-based approaches are better suited to capturing semantic and structural relationships between users. Multi-granular

Embedding Models have also been developed to increase the alignment and similarity between users on different social media platforms [17].

The major strength of the embedding-based approaches is that they allow combining various types of features into one representation space. In general, they provide better performance in recommendation systems, user profiling, and community detection problems. However, they are computationally costly and require significant amounts of data and computational power. Furthermore, it is difficult to interpret the learned representations, making the similarity reasoning process less intuitive than in other approaches.

Table 2.1: Comparative Summary of User Similarity Approaches on Twitter

Approach Type	Common Techniques	Main Features	Advantages	Limitations
Content-Based	TF-IDF, cosine similarity, LDA, hashtag analysis [18, 4]	Tweets, hashtags, profile descriptions	Simple, interpretable, effective for interest-based grouping	Affected by noisy and sparse textual data; ignores network structure
Graph-Based	Community detection, graph similarity, weakly supervised graph learning [3, 7]	Followers, followings, mentions, retweets, network connections	Captures structural relationships and interaction patterns	Computationally expensive; limited semantic understanding
Signal-Based	Weighted signal modeling, Random Forest, SVM, Naïve Bayes, k-NN [1, 5, 8]	Retweets, mentions, hashtags, likes, profile attributes	Flexible and effective for multi-dimensional similarity estimation	Dependent on feature selection and manual weighting strategies
Embedding-Based	Node2Vec, graph embeddings, multi-view representation learning, graph neural networks [6, 15]	Textual, structural, temporal, and behavioral features	Captures latent semantic and structural relationships with high performance	Computationally intensive with limited interpretability

2.4 Datasets and Features Used

The performance of user similarity analysis on Twitter is influenced by the type of dataset and features employed in modeling the user behaviors and interactions. Most works employ data obtained via Twitter API that comprises tweets, user profile details,

users' followers, followings, mentions, retweets, and other forms of interactions [1, 8]. In line with the applications, labeled and unlabeled datasets have been employed for supervised and unsupervised similarity modeling. Table 2.2 shows some user similarity works and their respective datasets.

Various approaches have incorporated a mixture of textual, network-based, interaction-based, temporal, and behavioral features for measuring similarity. Contemporary methods have integrated various types of heterogeneous features from different data sources for constructing user representation.

Table 2.2: Datasets Used in Selected Twitter User Similarity Studies

S.No.	Dataset Source	Dataset Characteristics
1	Real-time Twitter API dataset [4]	Contains user profiles, tweets, followers, retweets, favorites, mentions, and hashtag interactions collected from real Twitter users
2	Twitter ego-network dataset [18]	Includes ego users, friends, followers, interaction history, activity records, hashtags, and linguistic tweet content from interconnected Twitter networks
3	Twitter social graph dataset [3]	Composed of formal and inferred social connections, tweet corpora, interaction frequencies, and user-pair relationships
4	Temporal Twitter activity dataset [15]	Includes time-evolving tweets, user activities, streaming textual data, and temporal behavioral information collected over multiple time periods
5	Large-scale Twitter API dataset [1]	Contains seven behavioral and interaction signals including followers, followings, mentions, retweets, favorites, hashtags, and profile similarity
6	Semantic Twitter entity dataset [5]	Contains hashtags, mentions, replies, URLs, domains, and semantic relationships extracted from tweet content
7	Twitter timeline dataset of 19,900 users [8]	Includes behavioral ratios, audience networks, tweet timelines, interaction patterns, and textual content collected through the Twitter API
8	Multi-view social media dataset [6]	Contains timestamped posts, textual features, temporal activities, profile metadata, and network interactions across heterogeneous user populations

2.5 Evaluation Metrics

Evaluation metrics have been used extensively to study the performance of user similarity techniques on Twitter. The metric depends on how the problem formulation is approached, such as classification, ranking, clustering, or similarity estimation. Some

commonly used metrics clustering accuracy, Pearson correlation, and similarity score evaluation [1, 8]. Table 2.3 shows some of the metrics used and the main findings in selected researches.

Based on the results presented in the reviewed literature, the hybrid and multi-feature approaches show better performance than the single feature approaches since they incorporate all kinds of information such as textual, structural, and behavioral.

Table 2.3: Evaluation Metrics and Performance Summary of Selected Studies

S.No.	Evaluation Metrics	Approach Used	Key Findings
1	Precision, Recall, F1-score	Multi-signal similarity validation [1]	Improved similarity detection using combined behavioral signals
2	Accuracy, Precision, Recall, F1-score	Hybrid Random Forest model [8]	Hybrid features achieved high similarity prediction performance
3	Accuracy, Precision, Recall	SVM, MLP, k-NN, Naïve Bayes [4]	Machine learning models outperformed traditional similarity methods
4	Pearson correlation, similarity score	Interaction-based statistical analysis [18]	Interaction-based similarity was stronger than content-only similarity
5	ROC-AUC, Precision, Recall	Representation learning and ranking [6]	Multi-view embeddings improved ranking and representation quality
6	Accuracy, F1-score	Graph and community-based classification [3]	Structural network features were effective for user grouping
7	Clustering accuracy, similarity evaluation	Unsupervised clustering methods [19]	Effective grouping achieved using profile and textual features
8	Similarity score evaluation	Temporal embedding model [15]	Temporal embeddings captured evolving user behavior effectively
9	Accuracy, F1-score	Weakly supervised graph embedding [7]	Strong performance achieved with limited labeled data

2.6 Challenges, Research Trends, and Research Gaps

However, there are still a number of issues that need to be addressed despite the advances in the similarity assessment of Twitter users. For example, the sparsity of user-generated textual and interaction data can make it very difficult to estimate the similarity among users. Also, there is an inherent problem with the dynamic nature of user behavior, as interests and interactions change over time, but most methods use static representations [15]. Finally, scalability is another crucial issue, since large-

scale social networks typically comprise tens or hundreds of thousands of users and interactions.

It can be observed that the recent research directions are mainly focused on shifting from conventional content-based approaches to the hybrid or representation learning methods. In particular, modern methods usually incorporate textual, structural, behavioral, and temporal information into rich user representations [6, 8]. For instance, recent studies have paid much attention to graph embeddings, Node2Vec approach, graph neural networks, and multi-view representation learning. Recently, embedding distance techniques and graph autoencoders for studying echo chambers, polarizations, and ideological similarity in social media networks has been explored [20]. Distributed graph embedding models are other types of models that are gaining popularity when it comes to working with large-scale social networks comprising millions of nodes and edges [21]. Some recent work has also examined domain-specific embeddings to distinguish between different communities and polarized users within those communities [22].

Despite the promising results achieved by most of the approaches, some research gaps are currently evident in this area. First of all, there is no commonly accepted dataset for benchmarking the performance of different methods, and no unified framework for evaluation. In addition, many deep representation learning approaches face problems with interpretability and high computational requirements. Finally, there is not enough research dedicated to adaptive and real-time modeling of user similarities in dynamically changing social networks.

2.7 Summary

This chapter gives an overview of existing approaches for evaluating the user similarity measure on Twitter. Approaches are divided into four main categories, namely, content-based, graph-based, signal-based, and embedding-based approaches. Analysis was performed based on methodology, popular datasets used in the studies, common features, evaluation measures, strengths, and weaknesses of the approaches.

It can be seen from the overview provided that there is a trend towards using more complex approaches combining textual information with other types of information. Specifically, recent approaches consider textual, structural, behavioral, and temporal features for building better representations of users. At the same time, several key issues in the area have been identified, such as data sparsity, scalability, dynamics of user behavior, lack of interpretability, and benchmark datasets.

In general, it can be stated that no single method is able to evaluate user similarity

efficiently. Multi-view approaches show better results since they incorporate various types of information. Insights gained from the literature review are critical for understanding the comparative analysis performed in the next chapter.

Chapter 3

COMPARATIVE ANALYSIS OF USER SIMILARITY APPROACHES ON TWITTER

This chapter presents a comparative experimental analysis of three categories of similarity approaches: feature-based methods, representation learning-based methods, and metadata-driven methods.

The comparative study focuses on analyzing how effectively different approaches capture semantic, behavioral, and structural similarities among Twitter users. Experimental results are presented through quantitative evaluation and comparative analysis to provide a direct comparison between the implemented similarity models.

3.1 Dataset Collection

The comparative analysis is conducted based on Twitter data obtained from Twitter API-based and publicly available datasets. Datasets are chosen in order to assess the similarity of Twitter users in terms of semantics, behavior, structure, and metadata.

For the feature-based and representation learning-based techniques, Twitter data obtained from 19 active users is used. Data consists of tweets, hashtags, mentions, replies, retweets, and user interactions. Pairs of Twitter users are manually annotated as similar and dissimilar considering commonalities in topics covered, behaviors, and interactions. Obtained data varies in terms of textual content, timing, and behavior providing opportunities for comparative analysis in terms of semantic similarity, temporal behavior, and network interactions.

For the metadata-based similarity technique, publicly available Kaggle dataset consisting of 17,410 tweets made by 112 users is used. Dataset includes tweet text, user profile characteristics, interaction metadata, and other data necessary for the similarity assessment based on metadata.

3.2 Similarity Approaches

The three approaches take different perspectives on measuring user similarity by utilizing textual data, behavioral actions, metadata, and network relations.

The details of the three similarity approaches studied in this paper are provided in Table 3.1.

Table 3.1: Comparative Summary of Implemented Similarity Approaches

Approach	Techniques Used	Similarity Modeling
Feature-Based Similarity	TF-IDF, cosine similarity, DTW, Jaccard similarity, Random Forest [8]	Semantic similarity using tweet content, behavioral similarity using temporal activities, and structural similarity using audience overlap and interaction patterns.
Representation Learning-Based Similarity	Sentence-BERT embeddings, temporal embedding modeling, multi-view learning [6]	Unified user embeddings generated from textual, behavioral, temporal, and profile-level metadata features.
Metadata-Driven Similarity	TF-IDF, cosine similarity, Jaccard similarity, graph centrality analysis [19]	Similarity estimation using profile attributes, textual content, metadata features, and network-based graph analysis.

3.3 Performance Metrics

Evaluation metrics include both classification-based and embedding-based techniques for the implemented similarity approaches. Manual labeling of users' similarity is performed based on similarities in terms of their interests, behaviors, and network interactions.

Classification-based metrics such as Accuracy, Precision, Recall, and F1-score are utilized for evaluating feature-based and metadata-driven techniques. Accuracy evaluates the overall accuracy of user similarity prediction, whereas Precision is used to assess the ratio of the correct predictions of similar users to the total predictions of similar users. Recall evaluates the ability of the model to predict similar users accurately, and F1-score calculates the harmonic mean of Precision and Recall.

In addition to the above metrics, representation learning is evaluated using Cosine Similarity Analysis and Silhouette Score. Cosine Similarity Analysis evaluates the similarity of user embeddings generated by the model, whereas Silhouette Score evaluates the quality of clustering of user representations.

3.4 Results and Comparative Analysis

This section describes the results of the experiments conducted with respect to the feature-based, representation learning-based, and metadata-based approaches with respect to the classification and embedding metrics for evaluating similarity.

3.4.1 Feature-Based Similarity Results

The feature-based similarity model evaluates the similarity based on semantic, behavioral, audience features. The computation of the semantic similarity value using TF-IDF and cosine similarity ranges from 0.28 to 0.81, implying different degrees of textual similarity. The value obtained for the behavioral similarity measure using DTW is within the range of 30 to 50. The computation of the audience similarity measure using Jaccard similarity is quite low, with values from 0.00 to 0.11.

The Random Forest classifier demonstrates an accuracy of 0.60 along with precision, recall, and F1 score of 0.80, 0.60, and 0.57, respectively. The findings reveal semantic similarity measure offers consistent similarity signals, whereas behavioral and audience features add additional interaction information. However, the relatively poor classification performance may be attributed to the small dataset size and sparseness of interaction data [8].

3.4.2 Representation Learning Results

Representation learning techniques produce unified embeddings of users based on textual, behavioral, and profile-level information. The use of sentence-BERT embeddings together with temporal modeling based on embeddings enables capturing semantic and behavioral connections between users.

Cosine similarity between embeddings of similar users is over 0.95, demonstrating good embedding consistency. Embeddings training reveals that the loss function is reduced from 4.2 to 2.1, and the Silhouette score of about 0.68 indicates good clustering quality of learned embeddings. Also, the model manages to obtain precision, recall, and F1-scores of 0.85, 0.74, and 0.79 correspondingly for similarity classification based on embeddings.

The technique require more high computing resources and larger datasets to improve generalization performance [6].

3.4.3 Metadata-Driven Similarity Results

The metadata-based similarity method calculates similarities based on profile features, textual data, locational data, and metadata regarding interactions. The similarities calculated are between 0.45 and 0.57, showing that there is high similarity among the users.

The classification model has achieved an accuracy of about 0.88, with precision, recall, and F1-score values being 0.85, 0.79, and 0.82, respectively, performing better than other methods in terms of classification accuracy. Another method used for classifying users is graph-based centrality, which utilizes betweenness centrality measures to identify influential users in the network.

The metadata-based approach can be used to categorize users according to activities based on metadata and interactions in the network. This shows that the inclusion of both profile features and metadata increases classification effectiveness [19].

3.4.4 Comparative Analysis of All Approaches

From the above experiments, we can infer that user similarity in social networking sites is multi-dimensional in nature and cannot be accurately modeled using any one similarity approach. While the feature-based approach has greater interpretability by virtue of using explicit similarity features, the representation learning approach produces better embedding representations for representing user similarity.

Of all approaches tested, the similarity model based on metadata has yielded the best classification accuracy because of its ability to incorporate both profile attributes and interaction metadata. From the results obtained, it can be seen that each approach has its strengths and weaknesses, which make them suitable for different applications.

Table 3.2: Comparison of Different Similarity Approaches

Approach	Key Metrics	Key Strength	Limitation
Feature-Based	Accuracy: 0.60, F1-score: 0.57	High interpretability	Sensitive to noisy and sparse data
Representation Learning	Precision: 0.85, F1-score: 0.79	Captures latent semantic patterns	Lower interpretability
Metadata-Driven	Accuracy: 0.88, F1-score: 0.82	Better overall classification performance	Dependent on metadata quality

3.5 Discussion

It should be concluded that user similarity on Twitter cannot be measured by a single method and is multi-dimensional. All implemented approaches measure different aspects of users' activity, such as semantic relationships, temporal interaction patterns, profile attributes, and network structures.

Feature-based similarity model is more interpretable since similarity is calculated based on explicit semantic, behavioral, and structural features. The use of TF-IDF and cosine similarity helps to estimate semantic relationships between users. The use of DTW allows considering temporal behavioral patterns. At the same time, the overall classification accuracy remains modest because of the limitations of the dataset size and sparsity of interactions among users.

Representation learning model is more successful in capturing hidden semantic and temporal patterns through embedding-based modeling. The use of sentence-BERT embedding and temporal behavior encoding results in obtaining more compact and semantically consistent user representation. The estimated cosine similarity and silhouette score prove that embedding-based approaches are quite effective in revealing hidden similarity relationships among users. However, the approach requires bigger data and lacks interpretability in comparison with the previous method.

As for the metadata-driven similarity model, it turns out to be the most successful one when considering overall classification performance.

Therefore, each similarity paradigm has its unique strengths, which makes it suitable for various applications. Specifically, the first two methods are more interpretable than the last one, but the last one is more accurate. Combining different paradigms into one system may help to enhance future approaches.

One limitation of the study is the small dataset size used for evaluating the last two similarity paradigms. Furthermore, manual labeling of data introduces subjectivity and limits the generalizability of the models.

3.6 Summary

In this chapter, we have compared various methods of measuring similarity between users on Twitter, which include the feature-based method, representation learning-based method, and metadata-based method. It was found that each of these methods provides an insight into various dimensions of user similarity, such as semantic, behavioral, temporal, and network-related dimensions.

While the feature-based approach offers greater interpretability, the representation

learning approach captures complex latent similarities, and the metadata-based approach provides the best classification accuracy. This proves that user similarity on social networking sites is inherently multidimensional, and no single similarity measure can provide adequate results.

Overall, the experiment in this chapter has provided us with a comparative understanding of various similarity measurement methods.

Chapter 4

CONCLUSION, LIMITATIONS AND FUTURE SCOPE

4.1 Conclusion

Specifically, the feature-based model used explicit semantic, behavioral, and structural similarity measures, while the representation learning method created multi-view embeddings using Sentence-BERT and temporal embedding modeling techniques. The metadata-driven similarity technique included profile attributes, interaction metadata, and graph-based analysis in the similarity calculation process.

Based on the experimental results, it was concluded that user similarity in social media platforms is multi-dimensional and requires employing multiple similarity techniques for its accurate capture. While the feature-based approach yielded better interpretability due to explicitly computed similarity measures, the representation learning-based method was able to capture latent semantic relations. The best classification results were produced by the metadata-driven approach.

Overall, the study showed that hybrid and multi-view user similarity frameworks allow more efficient estimation of user similarity than single-source approaches. The research results are expected to help researchers gain a better understanding of modern Twitter user similarity techniques and conduct further comparative studies in this area.

4.2 Limitations

While the paper presents a comparative evaluation of various methods used to assess Twitter user similarities, some weaknesses can be identified. First, the experimental evaluation was performed based on the sample of a limited number of users whose similarity was assessed manually for both feature-based and representation learning methods. Such a small sample size may limit the generalizability of the results.

Second, the manual assignment of similarity labels may cause problems since user similarity itself is subjective and multidimensional, and therefore, manually labeling user similarities may cause certain discrepancies in assessing similarity.

Third, this research only considered textual, behavioral, profile-level, and network-based data available from Twitter. Other multimodal information, including images, videos, and geographic information, was ignored in similarity estimation.

Fourth, representation learning methods are usually computationally intensive. Finally, there are no benchmark datasets for evaluating Twitter user similarities that would allow comparing results obtained by other researchers.

4.3 Future Scope

Several possible directions for future research emerge from the results of this paper. First, the creation of hybrid or unified similarity frameworks, which combine feature-based, representation learning-based, and metadata-driven similarity methods in one system, should be considered as an important area for further research. Such integration of different similarity paradigms might increase both their interpretability and accuracy.

Future research might also be directed at the use of larger and more diverse datasets that will make the generalization capabilities of similarity models higher. Multilingual data, user interaction data that changes over time, and even information about social network interactions across various platforms may contribute to increased efficiency of similarity estimation.

Bibliography

- [1] Hind AlMahmoud and Shurug AlKhalifa. Tsim: a system for discovering similar users on twitter. *Journal of Big Data*, 5(1):39, 2018.
- [2] Ashish Goel, Aneesh Sharma, Dong Wang, and Zhijun Yin. Discovering similar users on twitter. In *11th Workshop on Mining and Learning with Graphs*, 2013.
- [3] John Conroy, Josephine Griffith, and Colm O’Riordan. Measuring twitter user similarity as a function of strength of ties. In *KDIR*, pages 262–270, 2011.
- [4] Ali Choumane and Fatima Yassin. Characterizing and detecting similar twitter users. In *2021 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*, pages 25–30. IEEE, 2021.
- [5] Gerasimos Razis and Ioannis Anagnostopoulos. Discovering similar twitter accounts using semantics. *Engineering Applications of Artificial Intelligence*, 51:37–49, 2016.
- [6] Siyi Guo, Keith Burghardt, Valeria Pant’e, and Kristina Lerman. Somer: Multi-view user representation learning for social media. *arXiv preprint arXiv:2405.05275*, 2024.
- [7] Tunazzina Islam and Dan Goldwasser. Twitter user representation using weakly supervised graph embedding. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 358–369, 2022.
- [8] Niloufar Shoeibi, Nastaran Shoeibi, Pablo Chamoso, Zakieh Alizadehsani, and Juan Manuel Corchado. A hybrid model for the measurement of the similarity between twitter profiles. *Sustainability*, 14(9):4909, 2022.
- [9] Nisha P Shetty, Balachandra Muniyal, Leander Melroy Maben, Rithika Jayaraj, and Sameer Saxena. Dynamic twitter friend grouping based on similarity, interaction, and trust to account for ever-evolving relationships. *IET Communications*, 18(17):1018–1048, 2024.

- [10] Sashank Sridhar and Sowmya Sanagavarapu. Twitter social networking graph using hypernym based semantic similarity detection. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pages 28–35. IEEE, 2020.
- [11] Riju Bhattacharya, Naresh Kumar Nagwani, and Sarsij Tripathi. A community detection model using node embedding approach and graph convolutional network with clustering technique. *Decision Analytics Journal*, 9:100362, 2023.
- [12] Miguel Won and Jorge M Fernandes. Analyzing twitter networks using graph embeddings: an application to the british case. *Journal of Computational Social Science*, 5(1):253–263, 2022.
- [13] Nir Lotan. *SocialVec: Social Entity Embeddings*. University of Haifa (Israel), 2021.
- [14] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.
- [15] Shangsong Liang, Xiangliang Zhang, Zhaochun Ren, and Evangelos Kanoulas. Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pages 1764–1773, 2018.
- [16] Ahmed El-Kishky, Thomas Markovich, Serim Park, et al. Twihin: Embedding the twitter heterogeneous information network for personalized recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2842–2850, 2022.
- [17] Yongqiang Peng, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xu Gu, and Peng Lu. Multi-granularity attribute similarity model for user alignment across social platforms under pre-aligned data sparsity. *Information Processing Management*, 61(6):103866, 2024.
- [18] Masoud Fatemi, Kostiantyn Kucher, Mikko Laitinen, and Pasi Fr"anti. Self-similarity of twitter users. In *2021 Swedish Workshop on Data Science (SweDS)*, pages 1–7. IEEE, 2021.
- [19] Andrea Tundis, Archit Jain, Gaurav Bhatia, and Max Muhlhauser. Similarity analysis of criminals on social networks: An example on twitter. In *2019 28th*

International Conference on Computer Communication and Networks (ICCCN), pages 1–9. IEEE, 2019.

- [20] Faisal Alatawi, Paras Sheth, and Huan Liu. Quantifying the echo chamber effect: An embedding distance-based approach. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 38–45, 2023.
- [21] Peng Fang, Arijit Khan, Siqiang Luo, Fang Wang, Dan Feng, Zhenli Li, Wei Yin, and Yuchao Cao. Distributed graph embedding with information-oriented random walks. *arXiv preprint arXiv:2303.15702*, 2023.
- [22] Wentao Xu and Kazutoshi Sasahara. Domain-based user embedding for competing events on social media. *Journal of Computational Social Science*, 9(1):15, 2026.