Social Media WebApp “TweetSphere”

Er. Natali Singla
CSE department
Chandigarh University
Mohali, Punjab, 140413, INDIA
e16482@cuchd.in

Kunal
CSE department
Chandigarh University
Mohali, Punjab, 140413, INDIA
21bcs3149@cuchd.in

Nishant Kumar
CSE department
Chandigarh University
Mohali, Punjab, 140413, INDIA
21bcs3325@cuchd.in

Dron Madaan
CSE department
Chandigarh University
Mohali, Punjab, 140413, INDIA
21bcs3365@cuchd.in

Ritesh Kumar
CSE department
Chandigarh University
Mohali, Punjab, 140413, INDIA
21bcs2212@cuchd.in

Krish
CSE department
Chandigarh University
Mohali, Punjab, 140413, INDIA
21bcs2056@cuchd.in

Abstract -The paper presents a novel approach to addressing the challenges of multiuser operations in Twitter by focusing on the expense of queries and the existing solutions for data partitioning. It introduces a method for reducing server interaction by partitioning data based on user interactions and implementing selective replication for frequently requested user data. This approach significantly improves partition quality, especially with low replication ratios. Additionally, the paper delves into the issue of Twitter spammers and their evolving tactics, proposing new detection features to combat spam accounts. It also explores the detection of duplicate fake accounts on Twitter using feature based analysis and machine learning techniques, with SVM demonstrating the best performance at 93.3% accuracy

I. Introduction

The challenges faced by social networks, particularly focusing on the issues related to the availability and scalability of these platforms. The growth and dynamic nature of social networks demand high availability, which is addressed, in part, by employing NoSQL systems. These systems utilize data partitioning and replication, specifically hash-based partitioning and random replication, to achieve scalability and availability. However, the text points out that the current approach may lead to redundant replication, resulting in a significant number of writes on replicas and subsequent performance degradation.

To cope with the large volume of data and massive read and write requests, companies often maintain clusters with thousands of commodity hardware machines. Traditional relational databases are deemed unsuitable for this domain due to the negative impact of joins and locks on distributed system performance. Additionally, the text highlights the importance of high availability in addition to high performance, necessitating databases to provide failover mechanisms and easy replicability.

The discussion then shifts to online social networking websites, with a specific focus on Twitter, one of the most popular and fastest-growing platforms with over 190 million accounts tweeting 65 million times a day. The text addresses the challenges posed by malicious activities on Twitter, such as spam, malware. Distribution, and other illicit activities. Spammers on Twitter have evolved to evade detection mechanisms, leading to an ongoing "arms race" between security researchers and spammers. The paper aims to design more robust features to detect Twitter spammers by analysing evasion tactics employed by current spammers. The authors collected and analysed a substantial dataset of around 500,000 Twitter accounts and more than 14 million tweets. They identified around 2,000 Twitter spammers using blacklist and honeypot techniques, aiming to understand and validate evasion tactics through case studies and evaluating existing state-of-the-art approaches.

The next part of the text delves into the nature of Twitter usage, emphasizing how users share their feelings, news, events, and daily activities. It discusses how malicious users exploit Twitter to spread false stories, links, and images, prompting the need to identify false accounts on the platform. The narrative underscores the importance of detecting fake accounts for users' safety.

The text then broadens its scope to social networking sites in general, such as Facebook, Twitter, and Instagram, becoming leading platforms for communication and expression. Twitter, in particular, is highlighted as a popular service with 330 million active users sending 500 million tweets daily. The text acknowledges the attraction of spammers to Twitter for spreading hateful URLs, rumours, and unsolicited messages.

Another perspective introduced in the text revolves around the rise of social media platforms in the last decade. With the exponential increase in traffic on these platforms, issues such as developing duplicate accounts have emerged. The text mentions the creation of spam environments by duplicate accounts, especially those impersonating celebrities, high officials, and influencers.

It proposes feature-based analysis to detect duplicate fake accounts, employing techniques like decision tree, random forest, and SVM, with SVM performing better than other machine learning techniques.

II. Literature Survey

This section will delve into the historical evolution of social media platforms, tracing the development from early predecessors to contemporary platforms like Twitter. It will analyze key milestones, innovations, and user adoption trends shaping the landscape of social networking.

2. Real-Time Communication Technologies:

This segment will explore the underlying technologies enabling real-time communication in web applications, with a focus on their implementation in social media platforms. It will review WebSocket, Server-Sent Events (SSE), and similar protocols, assessing their suitability for providing instant updates and notifications in TweetSphere.

3. User Engagement and Interaction:

This section will examine research on user engagement metrics and interaction patterns in social media platforms. It will investigate the impact of features like likes, comments, retweets, and hashtags on user behavior and content virality, providing insights to inform the design of engaging features in TweetSphere.

4. Privacy and Security in Social Media Platforms:

This part will analyze the privacy and security challenges faced by social media platforms, including data privacy concerns, identity theft, and user tracking. It will explore strategies for enhancing privacy protection and security measures in web applications, offering recommendations for ensuring user trust and confidence in TweetSphere.

5. Content Moderation and Community Guidelines:

This segment will focus on content moderation practices and community guidelines in social media platforms. It will discuss the challenges of detecting and mitigating harmful content such as hate speech, fake news, and abusive behavior, proposing approaches for effective moderation and fostering a positive online environment in TweetSphere.

6. User Experience Design and Interface:

This section will review user experience design principles and best practices for designing intuitive and engaging interfaces in web applications. It will explore research on interface design patterns, accessibility considerations, and mobile responsiveness, guiding the development of an immersive user experience in TweetSphere.

II. Design and implementation

.

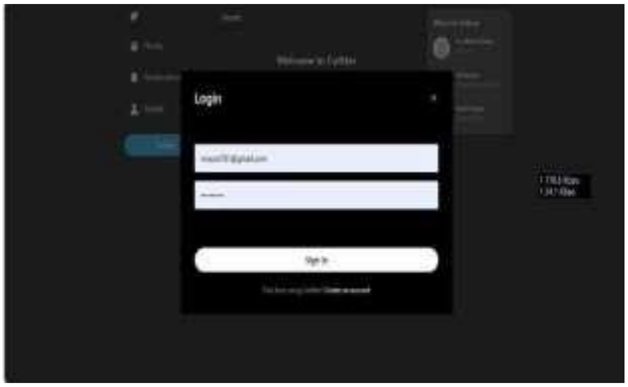


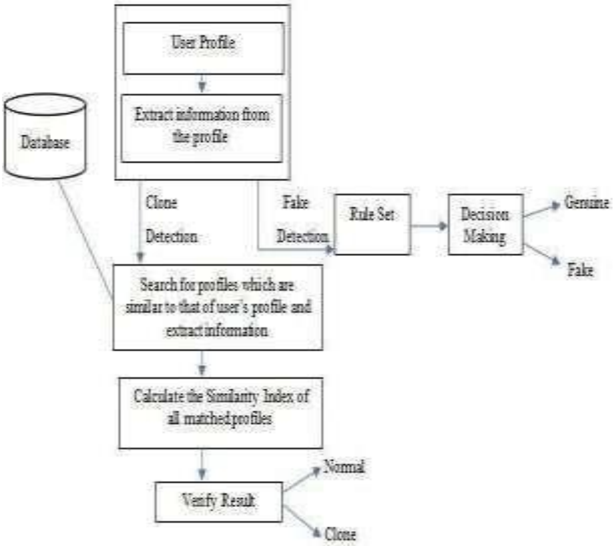
Fig. 1: Overview of Authentication

In the implementation of the Twitter clone, Twissandra, the emphasis is on leveraging the Cassandra NoSQL database to manage the six key column families: USER, FOLLOWINGS, FOLLOWERS, TWEET, HOMETIMELINE, and USERLINE. This database design facilitates the basic functionalities inherent in Twitter, accommodating the complexities that arise during operations involving multiple users. Particularly, when a user posts a tweet, the system is engineered to handle the propagation of the tweet to the user's followers, necessitating multi-write operations. Similarly, when a user checks their homepage to load the latest tweets of those they follow, the system is optimized for multi-read operations.

The distinction between operations on the same or different servers is acknowledged, underlining the scalability considerations in the system's architecture. To address the persistent challenge of spammers on Twitter, the text introduces a method focused on developing robust detection features. These features are strategically designed to be either difficult or expensive for malicious entities to evade, thereby enhancing the platform's security. Among the proposed features are graph-based features and neighbour- based features, leveraging the inherent structure of the Twitter network. Key features include the local clustering coefficient, which measures the degree of interconnectedness among a user's immediate connections; betweenness centrality, which identifies users who act as bridges between different segments of the network; and bidirectional links ratio, which assesses the proportion of reciprocal connections a user has. These features collectively contribute to a comprehensive strategy for identifying and distinguishing spammers by analysing their social behaviours.

The text specifically delves into the domain of fake profile detection, elucidating rules that serve as criteria for distinguishing between genuine and fake profiles. The absence of a profile name or image, lack of account description, a false geo-enabled field, and an unusually substantial number of tweets are highlighted as indicators of potential fake profiles. Notably, the emphasis is placed on the implementation of features that necessitate fundamental changes in the behaviour of malicious entities, aiming to counter their evasion tactics effectively. Another facet of the proposed method involves clone profile detection, where the text introduces the application of similarity measures based on both attributes and network structures. Attribute similarity, assessed through metrics such as cosine similarity and Levenshtein distance, scrutinizes attributes like name, screen name, language, location, and zone. This method is designed to identify profiles that closely resemble a user's profile, indicating potential clones. Additionally, the C4.5 algorithm, a decision tree algorithm, is introduced for clone profile detection. This algorithm builds decision trees based on provided data, with attributes such as information gain and entropy guiding the selection of the most effective attributes for decision-making. The use of the C4.5 algorithm in this context underscores a systematic approach to distinguishing between genuine and cloned profiles.

The text further details the intricacies of fake profile detection, emphasizing the significance of features that necessitate substantial changes in malicious entities' behaviours. The introduction of clone profile detection, employing similarity measures and decision tree algorithms, adds another layer of sophistication to the platform's security measures.


User Profile
Extract information from the profile
Database
Clone Detection
Fake Detection
Rule Set
Decision Making
Genuine
Fake
Search for profiles which are similar to that of user's profile and extract information
Calculate the Similarity Index of all matched profiles
Verify Result
Normal
Clone

III. **Data Collection And Clone Identification**

Clone Set Identification Framework: To collect clone sets, we developed a framework consisting of three main components: (1) a tweet retriever, (2) a text extractor, and (3) a clone finder. First, the tweet retriever retrieves tweets containing links to news articles using Twitter’s API Academic Researcher product track. Here, we first obtained all tweets posted within an example timeline that contained an URL and then filtered the (resolved) URLs against the domains owned by a list of the most popular US news outlets. Second, the text extractor was used to extract the news article texts (but not figures, videos, etc.) from the URLs. Here, we used a combination of the opensource news-please Python module and a custom text extractor that we implemented for news websites with more complex structures (that performed poorly on), as well as a per- domain specific crawler. Our custombuilt crawler was built using the library Beautiful Soup. Finally, the clone finder module identifies potential clones by first grouping all tweets using the same URL and then applying a two-phase clone (or near clone) identification approach on the extracted texts (when available). With the two-phase identification, we first use Simhash [4], a technique that generates 64-bit fingerprints for each text (64- bit has helped avoid collisions compared to 32-bit hashes), to check for similarity using a maximum hamming distance of 6 (ensuring a high recall), followed by calculating the pairwise cosine similarity on the vectorized texts (created using TF-IDF) of all pairs within a candidate cluster, so as to further refine the clone sets (and improve the precision). Our manual inspection showed that using a similarity threshold of 0.9 and combining these two phases (i.e., for initial candidate clone identification followed by pairwise similarity tests within a cluster) reduces computational complexity (i.e., limits the required pairwise tests) and enhances accuracy (i.e., avoids unnecessary exclusion of potential clones, enabling more rigorous assessment in the subsequent cosine similarity). News Outlet Selection and Data Preparation: To select news outlets (for URL filtering and text extraction), we used the ranking lists of several independent rankings of US news outlets (e.g., opensources.co,. The 69 selected news outlets represent a diverse range of topics, geographical locations, and audiences. Datasets: Four datasets were collected based on the age of each post at the time of data collection (one- year old, half a-year old, one-month old, and one week old) and for each dataset we collected two snapshots: one when the posts are of the above listed ages and one that was collected one week later. In both cases, we collected all possible statistics about the tweet (including retweet statistics) and the tweeter of the tweet. The different aged datasets allowed us to analyse the effect of age differences on retweet behaviour, while the retweet recollection one week later allows us to evaluate and reflect on the stability of the retweet counts over time. Table I summarizes the size of the datasets. All datasets were collected over the week of March 2-8, 2021. Combined, the four datasets consist of 4.5M unique tweets including links to one of our predetermined URLs. Of these, most tweets (3.9M) are part of one of the 274,510 identified clone sets. Success metric: To measure the successful spread of a tweet, we primarily use the number of retweets. This choice is motivated by the high importance of recommendations by friends and family (e.g., 83% believe more in such trust-earned advertisements than regular advertisements [5]) and world advertisement in general. We also show results for other public interaction metrics such as likes, quotes, and replies.

Age	Total Tweets	Tweets In Clone sets	Total Clone Sets	Largest Clone Sets
1 year	1,398,359	1,219,244	75,902	10,165
½ year	1,128,696	988,331	70,773	4,057
1 month	1,021,421	883,816	65,151	6,673
1 year	928,587	811,558	62,684	9,146

Table I Dataset Overview

I

.

IV. DISCUSSION

social networks, particularly Twitter, in terms of scalability, availability, and performance. It highlights the use of NoSQL systems to address these challenges through data partitioning and replication. The dynamic nature of social networks, coupled with the need for high availability, drives the adoption of these distributed systems. However, the drawbacks of traditional relational databases, such as the impact of joins and locks in distributed systems, make them unsuitable for this context.

NoSQL systems employ hash-based partitioning and random replication to distribute data across clusters of commodity hardware machines. While this approach helps achieve scalability and availability, it often results in redundant replication and a high volume of writes on replicas, leading to performance degradation. The need to respond to massive read and write requests without latency requires maintaining clusters with thousands of machines.

The limitations of relational databases in terms of replication techniques and their focus on consistency over availability make them less suitable for the demands of social networks. Data partitioning is introduced as a technique to improve query response times for I/O intensive applications. The goal is to optimize query processing time by reducing the number of servers required for a single user while answering a query. Replication of data items is also explored to achieve higher I/O parallelism, particularly in the context of optimizing range queries.

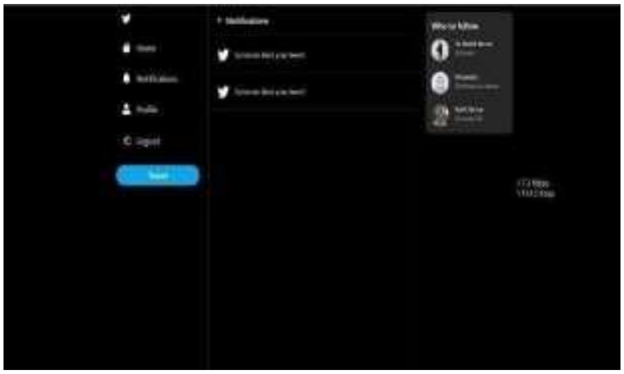
However, the passage acknowledges challenges associated with data replication, such as the need to consider consistency during update and delete operations, the potential slowdown of write operations, and the extra storage requirements. It emphasizes the importance of avoiding unnecessary replication, especially in applications with very large data sizes.

The primary focus of the paper is then introduced: the proposal of a workload-aware replication and partitioning method specifically tailored for Twitter.
The subsequent sections are outlined, including a review of related works on data partitioning and replication in social networks, a detailed presentation of the proposed model, experiments and results, and a concluding section.

The passage briefly touches upon related works in the field, highlighting studies that analyse the topological characteristics of Twitter accounts, metrics for measuring account behaviour, and techniques to prevent spam and attacks on social networks. It mentions existing approaches to detecting Twitter spammers, categorizing them into machine learningbased methods and those based on examining URLs and domains for malicious content. The paper's approach is positioned as addressing the evolving tactics of spammers, focusing on designing more robust features to detect evasive spammers and qualitatively analysing the robustness of detection features.

Several other papers are referenced briefly, covering topics such as the protection of Twitter trends from hateful users, trust relationships among social media users, reliability analysis to prevent the spread of false information on Twitter, and the influence of word calling trends on microblog creators. These works use various methodologies, including machine learning, to address issues such as spam detection, account legitimacy, and the impact of trends on user behaviour.

In summary, the passage provides an overview of the challenges faced by social networks, particularly Twitter, in terms of scalability, availability, and performance. It discusses the use of NoSQL systems, data partitioning, and replication to address these challenges. The paper's primary focus is on proposing a workload-aware replication and partitioning method for Twitter, with related works providing context and insights into various aspects of social network analysis and improvement.



Front-end design

V. CONCLUSIONS

In conclusion, the creation of a Twitter clone website presents a multifaceted endeavor that encompasses technical, ethical, and user experience considerations. This research paper delved into the intricacies of developing a platform mimicking the functionalities of Twitter. The benefits of undertaking such a project include providing a rich learning experience for developers, customization opportunities to tailor features to specific needs, and the potential for innovation in terms of design and technology integration.
Additionally, a successfully executed Twitter clone can serve as a valuable addition to a developer's portfolio, showcasing their skills and versatility.

However, the challenges associated with this endeavour are equally significant. Security concerns, including the safeguarding of user data and ensuring privacy through robust encryption and authentication measures, must be prioritized. Scalability is another critical aspect, necessitating careful architectural planning to accommodate potential growth in user traffic and data volume. Moreover, legal and ethical considerations loom large, requiring adherence to intellectual property rights, terms of service, and user privacy regulations. Striking the right balance between innovation and ethical responsibility is paramount for the project's long- term success.

As the social media landscape evolves, user engagement emerges as a pivotal challenge. Creating features that captivate users, responding to feedback, and adapting to shifting user expectations are ongoing concerns that demand continual attention and improvement. In the dynamic realm of social media, the success of a Twitter clone hinges on its ability to stay relevant, secure, and engaging.

In summary, while the development of a Twitter clone offers numerous advantages for developers, it demands a comprehensive and nuanced approach. Success is contingent on addressing the intricate interplay between technical, ethical, and user-centric considerations. As technology and user preferences evolve, the journey of refining and enhancing the Twitter clone becomes a perpetual pursuit for excellence in the ever-evolving landscape of social media platforms.

The passage addresses the growing challenges of fake and clone profiles on online social networks, particularly Twitter, proposing a robust detection method. Fake profiles are discerned using a set of rules, while clone detection employs Similarity Measures and the C4.5 algorithm, with Similarity Measures proving more effective in identifying significant portions of cloned profiles. The passage introduces a novel selectively replicated partitioning method for data partitioning and replication on Twitter, leveraging a temporal activity hypergraph to address drawbacks in existing approaches. Users undergo initial random partitioning into two groups, refined through recursive processes. The paper significantly contributes to Twitter spam detection by introducing effective features based on an in-depth analysis of spammers' evasion tactics. These features outperform existing detectors across machine learning classifiers, promoting further research through the release of related datasets.

VI. References

[1] Z. Miao and L. Fan,A novel multiagent decision making architecture based on dual's dual problem formulation, IEEE Transactions on Smart Grid, 99(2016) 1–1.

[2] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, Network anomaly detection: Methods, systems and tools, IEEE Commun. Surv. Tutor., 16(1)(2014) 303–336.

[3] M. Weiten, Ontostudio® as an ontology engineering environment,in Semantic knowledge management. Springer, (2009) 51–60.

[4] N. M. Meenachi and M. S. Baba,Web ontology language editors for the semantic web-a survey, International Journal of Computer Applications, 53(2012) 12.

[5] M. De Choudhury, Y.-R. Lin, H.
Sundaram, K.S. Candan, L. Xie, and A. Kelliher, "How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?" Proc. Fourth Int'l AAAI Conf. Weblogs and Social Media, 2010.

[6] Atodiresei, C. S., Tănăselea, A., &
Iftene, A. (2018). Identifying fake news and fake users on Twitter. Procedia Computer Science, 126, 451- 461.

[7]H. Kwak, C. Lee, H. Park, and S. Moon,
"What is Twitter, a social network or a News Media?," in Proc. Int. World Wide Web (WWW'10), Raleigh, NC, USA, 2010.

[8] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P.Markatos, "Detecting Social Network Profile Cloning", 2013
.

[9] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks, 61(2015) 85–117.

[10] R. Vijayakumar, M. Alazab, K. P. Soman, P.
Poornachandran, A. Al-Nemrat, and S. Venkatraman, Deep learning approach for an intelligent intrusion detection system, IEEE Access,7,(2019) 41525–41550.