ICIW 2023

The Nineteenth International Conference on Internet and Web Applications and Services

ISBN: 978-1-68558-147-3

April 14 - 18, 2024

Venice, Italy

ICIW 2024 Editors

Jaime Lloret Mauri, Universitat Politecnica de Valencia. Spain

ICIW 2024

Forward

The Nineteenth International Conference on Internet and Web Applications and Services (ICIW 2024), held between April 14th and April 18th, 2024, continued a series of international events that covered the
complementary aspects related to designing and deploying of applications based on IP&Web techniques and mechanisms.

Internet and Web-based technologies led to new frameworks, languages, mechanisms and protocols for Web applications design and development. Interaction between web-based applications and classical applications requires special interfaces and exposes various performance parameters.

Web Services and applications are supported by a myriad of platforms, technologies, and mechanisms for syntax (mostly XML-based) and semantics (Ontology, Semantic Web). Special Web Services based applications such as e-Commerce, e-Business, P2P, multimedia, and GRID enterprise-related, allow design flexibility and easy to develop new services. The challenges consist of service discovery, announcing, monitoring and management; on the other hand, trust, security, performance, and scalability are desirable metrics under exploration when designing such applications.

Entertainment systems became one of the most business-oriented and challenging areas of distributed real-time software applications and special devices industry. Developing entertainment systems and applications for a unique user or multiple users requires special platforms and network capabilities.

Particular traffic, QoS/SLA, reliability and high availability are some of the desired features of such systems. Real-time access raises problems of user identity, customized access, and navigation. Services such interactive television, car/train/flight games, music and system distribution, and sport entertainment led to ubiquitous systems. These systems use mobile, wearable devices, and wireless technologies.

Interactive game applications require methodologies, frameworks, platforms, tools and languages. State-of-the-art games today can embody the most sophisticated technology and the most fully developed applications of programming capabilities available in the public domain.

The impact on millions of users via the proliferation of peer-to-peer (P2P) file sharing networks such as eDonkey, Kazaa and Gnutella was rapidly increasing and seriously influencing business models (online services, cost control) and user behavior (download profile). An important fraction of the Internet traffic belongs to P2P applications.

P2P applications run in the background of user's PCs and enable individual users to act as downloaders, uploaders, file servers, etc. Designing and implementing P2P applications raise particular requirements. On the one hand, there are aspects of programming, data handling, and intensive computing applications; on the other hand, there are problems of special protocol features and networking, fault tolerance, quality of service, and application adaptability.

Additionally, P2P systems require special attention from the security point of view. Trust, reputation, copyrights, and intellectual property are also relevant for P2P applications. On-line communications

frameworks and mechanisms allow distribute the workload, share business process, and handle complex partner profiles. This requires protocols supporting interactivity and real-time metrics.

Collaborative systems based on online communications support collaborative groups and are based on the theory and formalisms for group interactions. Group synergy in cooperative networks includes online gambling, gaming, and children's groups, and at a larger scale, B2B and B2P cooperation.

Collaborative systems allow social networks to exist; within groups and between groups there are problems of privacy, identity, anonymity, trust, and confidentiality. Additionally, conflict, delegation, group selection, and communications costs in collaborative groups have to be monitored and managed. Building online social networks requires mechanism on popularity context, persuasion, as well as technologies, techniques, and platforms to support all these paradigms.

Also, the age of information and communication has revolutionized the way companies do business, especially in providing competitive and innovative services. Business processes not only integrate departments and subsidiaries of enterprises but also are extended across organizations and to interact with governments. On the other hand, wireless technologies and peer-to-peer networks enable ubiquitous access to services and information systems with scalability. This results in the removal of barriers to market expansion and new business opportunities as well as threats. In this new globalized and ubiquitous environment, it is of increasing importance to consider legal and social aspects in business activities and information systems that will provide some level of certainty. There is a broad spectrum of vertical domains where legal and social issues influence the design and development of information systems, such as web personalization and protection of users privacy in service provision, intellectual property rights protection when designing and implementing virtual works and multiplayer digital games, copyright protection in collaborative environments, automation of contracting and contract monitoring on the web, protection of privacy in location-based computing, etc.

We take here the opportunity to warmly thank all the members of the ICIW 2024 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to ICIW 2024. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the ICIW 2024 organizing committee for their help in handling the logistics of this event.

We hope that ICIW 2024 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of Internet and Web applications and services.

ICIW 2024 Chairs

ICIW 2024 Steering Committee
Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain
Petre Dini, IARIA, USA/EU
Dawid Mieloch, Poznan University of Technology, Poland

ICIW 2024 Publicity Chairs
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

ICIW 2024
Committee

ICIW 2024 Steering Committee

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain
Petre Dini, IARIA, USA/EU
Dawid Mieloch, Poznan University of Technology, Poland

ICIW 2024 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

ICIW 2024 Technical Program Committee

Mohd Helmy Abd Wahab, University Tun Hussein Onn Malaysia, Malaysia
Witold Abramowicz, Poznan University of Economics and Business, Poland
Santosh Nikhil Kumar Adireddy, ByteDance, USA
Samer Al-Khateeb, Creighton University, USA
Grigore Albeanu, Spiru Haret University, Bucharest, Romania
Nada Ayman M. Gab Allah , American University in Cairo / Coventry University, Egypt
Ado Adamou Abba Ari, University of Maroua, Cameroon
Suilen H. Alvarado, Universidade da Coruña (UDC), Spain
Leonidas Anthopoulos, University of Thessaly, Greece
Ezendu Ariwa, University of Bedfordshire, UK
Jocelyn Aubert, Luxembourg Institute of Science and Technology (LIST), Luxembourg
Chiara Bachechi, University of Modena and Reggio Emilia, Italy
Ismail Badache, LIS | Aix-Marseille University, France
Ibtissam Bakkouri, Ibn Zohr University, Agadir, Morocco
Masoud Barati, Cardiff University, UK
Yassine Ben Ayed, University of Sfax | Higher Institute of computer science and multimedia of Sfax, Tunisia
Maumita Bhattacharya, Charles Sturt University, Australia
Letizia Bollini, Free University of Bozen-Bolzano, Italy
Gabriela Bosetti, VeryConnect, Scotland
Christos Bouras, University of Patras, Greece
Tiago Brasileiro Araujo, Federal Institute of Paraíba, Brazil
Tharrenos Bratitsis, University of Western Macedonia, Greece
Francesco Buccafurri, Università degli Studi Mediterranea di Reggio Calabria, Italy
Anis Charfi, Carnegie Mellon University - Qatar Campus, Qatar
WaiShiang Cheah, University Malaysia Sarawak, Malaysia
Dickson Chiu, The University of Hong Kong, Hong Kong
Vincenzo De Angelis, University of Reggio Calabria, Italy
Luis M. de Campos, University of Granada, Spain
Francisco José Domínguez Mayo, University of Seville, Spain
Thomas Dreibholz, SimulaMet - Simula Metropolitan Centre for Digital Engineering, Norway

Sherali Zeadally, University of Kentucky, USA
Pengkai Zhu, Amazon Web Services (AWS) AI, USA
Jizhe Zhou, University of Macau, Macau, China

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Digital Uses and Practices in Secondary School

Achieving Digital Literacy

Hafida Hammadi
DICEN IDF – Gustave Eiffel University
Marne-la-Vallée, France
E-mail: hafi_12@yahoo.fr

Christian Bourret
DICEN IDF – Gustave Eiffel University
Marne-la-Vallée, France
E-mail: christian.bourret@univ-eiffel.fr

***Abstract*— Our study focuses on the description and analysis of the digital activities that teachers and students engage in both in-person and online at Georges Seurat Middle School in Courbevoie, in the region of Ile-de-France, as well as the digital content that is made available to students in face-to-face, distance learning, and hybrid courses. In this context, our qualitative research is based on an average of 314 hours of observation across a variety of digital projects. This includes 80 hours per academic year beginning in 2021. We sampled 397 students in grades seven through ten. In addition to the qualitative investigation, the quantitative study's findings are the main focus of this paper. We questioned 200 pupils, ranging in grade from seven to ten, regarding their use of and behaviors with digital devices in both private and academic settings.**

Keywords- digital literacy; digital workspace; digital practices; digital education; media literacy; virtual learning environment; virtual collaborative system.

I. INTRODUCTION

Our qualitative research was conducted during educational sessions in the school's digital collaborative space, OZE92. Several digital projects have been established to enhance media and information education, including a fake news project, an O-Lab citizen project, and a media class. Students collaborate online to finish projects under the guidance of the teacher. 314 hours of observation between September 2021 and March 2024 were used to obtain the results presented in this paper.

The qualitative study's findings indicate how group work in the context of online collaborative spaces -OZE92- engages students in an active learning process, motivates them to complete their digital projects [1][2], and helps students who are struggling to feel like they belong to the collaborative group.

Furthermore, our research shows how the role of the teacher has changed throughout time. Teachers are no longer the only ones who can impart knowledge; students are now actively involved in creating it. The supervisor's position has evolved from that of the teacher. This research is based on 314 hours of observations that were completed within OZE92's [3] collaboration spaces. In addition to the quantitative data that we reported in our previous work [14], this paper gives the findings of the qualitative investigation.

Taking group sample of pupils', teachers', and parents' digital behaviors and uses in both public and private settings is the aim of this study. This could provide us with an extensive knowledge of the requirements and expectations for education with the goal of gaining digital education [9] [10][12]. Consequently, is it feasible to discuss the development of digital literacy in higher education through analyzing the qualitative and quantitative findings of every member of the academic community? We have some responses to that question. The study's global analysis is currently in progress.

The rest of the paper is structured as follows. Section 2 covers the theoretical foundations of our research, Section 3 analyzes the qualitative study related to students, Section 4 discusses the limitations of our study, and Section 5 offers a partial conclusion of this research.

II. SCIENTIFIC POSITIONING

As stated in [11], our research highlights the principles of the "constructivist" approach to communication sciences and has connections to educational sciences.

The Anthony Giddens and Poole Scott Adaptive structuration theory serves as the foundation for this work [4]. AST theory focuses on the rules, resources, and social structures offered by institutions and technologies as the foundation of human behavior. Action frameworks and technology frameworks are always connected.

The authors of [5][6] and [7] present the idea of Personal Learning Environment (PLE), which puts the student at the center of the system. It is an ecosystem or collection of tools that supports learners in planning and organizing their learning.

Furthermore, Moore and Marty [8] were able to comprehend that learners need to be more autonomous depending on how extended the transactional distance is, according to the notion of transactional distance.

A transactional distance learning program needs to include a structure, an interaction, as well as a degree of autonomy for its students.

III. QUANTITATIVE RESULTS -STUDENTS

200 school students in grades seven through ten were asked about their digital practices and uses [12]. Students were questioned within the context of the quota technique [13]. The quota method is a sampling strategy in which distinct targets or quotas are set for different groups within a

 ISBN: 978-1-68558-147-3

population according to particular criteria. These quotas are intended to make sure that, in terms of those attributes, the final sample reflects the diversity of the population. In our case study, a sampling method was employed to select a subset of students for investigation from a total population of 600. Specifically, 200 students were chosen for participation. To ensure representation across various grade levels, 50 students were selected from each grade level, ranging from seven to ten grade.

Inside the private environment, 98% of students have access to the Internet at home (Figure 1) and are given the essential digital equipment. 35% of them have more than four smartphones available at home. 56% possess one tablet, 33% one laptop, 60% are connected to a game console, and 65% have desktop computers.

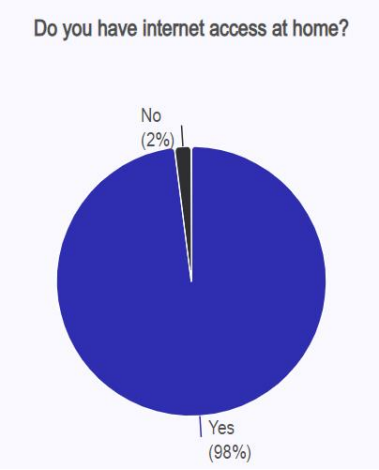

Figure 1. Internet access at home.

90% of students have a smartphone as their personal device, whereas 63% use a laptop. In their rooms, 79% of students set up their own digital devices.

They are connected daily (95% of the week), with 60% of them being online more than one hour per day. Students use the Internet primarily to chat with friends (73%), consult OZE92 (76%) and play online games (66%).

During the week, students consult the following sites in particular: YouTube (80%), OZE92 (78%) and social networks (75%). 93% of students use the Internet on Saturdays and Sundays. 65% of them stay online for more than an hour, mostly to watch videos on YouTube (79%), communicate with friends on social networks (75%), to play games online (67%) and, in fourth place, to consult OZE92 (65%). YouTube (80%), social networks (78%), and OZE92 (65%) are the most frequently visited websites on weekends.

86% of students use social media, particularly WhatsApp (92%), followed by Snapchat (78%), TikTok (76%), and Instagram (70%). Social networks are mostly utilized for interacting with friends (98%), keeping up with the news (62%), and watching videos on various topics (Figure 2).

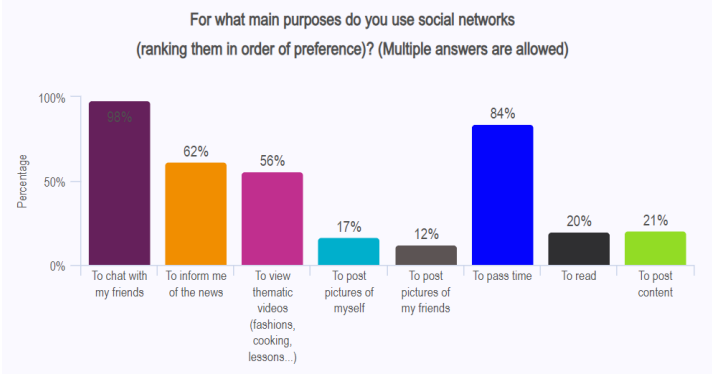

Figure 2. The purposes of social networks' uses.

In the classroom, the Internet is used primarily in technology (83%), mathematics (47%), sciences and English subjects (28%). Most of the digital equipment used in the classroom are ultra-portable computers (85%) and desktop computers (73%). Lessons (91%) and exercises (90%) constitute the majority of the digital content displayed in the classroom. Teachers often request digital outputs in the form of slideshows (87%), videos (38%), and audio files (35%). Most of the digital assignments that teachers give are document-based research (78%) and activities (70%).

Wikipedia is considered as a primary documentary source for carrying out research (91%), with YouTube following in the second (65%) and social networks in the third position (67%). Students use the OZE92 content for exercise resolution (76%) and course revision (72%). When it comes to documentary research, Wikipedia (82%) and OZE92 (65%) are still the most often utilized resources in the classroom for activities and revision courses, respectively. 51% of students use YouTube as a resource, placing it in a third place.

VI. STUDY'S LIMITATIONS

A few methodological issues with the selection of the student sample were present in our study. We would have preferred a random sample that was "statistically ideal" according to De Singly [13] as each member of the population surveyed must have an equal probability of being included in the sample. However, it was logistically unfeasible to present the survey to randomly selected students considering the constraints imposed by school scheduling.

V. CONCLUSIONS

According to our qualitative study, we have learned more about the digital profiles of teenagers attending Georges Seurat School. The data gathered shows that teenagers are well-equipped, especially in terms of smartphones (90%). They are daily connected (95%) on a week, 60% among them are online more than one hour per day. They frequently utilize the internet for both academic and personal purposes. They mostly use the internet for OZE92 consulting (76%), and social media conversation

 ISBN: 978-1-68558-147-3

with friends on networks (73%), with Instagram, TikTok, and Snapchat being the most popular platforms. Wikipedia and YouTube are the primary resources in the context of document-based research, both at home and in the classroom. Digital contents in the classroom specifically refer to lessons and activities that are projected onto a screen via a video projector. In this context, teachers specifically request digital research and activities.

These first quantitative findings may be applied to secondary schools that have students with similar characteristics (age, school level, social environment, etc.) and access to adequate digital resources both at home and at school. The response from the teacher questionnaire will be merged with the current data to determine whether the school has a well-established culture of digital literacy. These outcomes could be used, furthermore, by officials to enhance school programs and make greater use of digital technologies for education.

REFERENCES

[1] J. Boissière and E. Bruillard, "The digital school: a constructible and immersive learning environment", Ed. Armand Colin, 2021, p. 367.

[2] R. Hétier (Directed by), "Education's use of digital technology", Ed. Le bord de l'eau, 2021, p. 253.

[3] https://enc.hauts-de-seine.fr/my.policy. Ed. Ditrict Hauts-de-Seine, Ile-de-France, France, [retrieved: February, 2024].

[4] M. Bia Figueiredo and A. Adrot, "Gerardine DeSanctis and Marshall Scott Poole - The founders of adaptive structuring theory" in "Great Authors in Information Systems", 2018. pp. 291 – 309.

[5] G. Attwell, "Personal Learning Environments-the future of eLearning? https://www.researchgate.net/publication/228350341.Ed.researchgate.net. [retrieved: June, 2023].

[6] O. Liber and M. Johnson,"Personal_Learning_Environments". https://www.researchgate.net/publication/220095391.Ed.researchgate.net. [retrieved: June, 2023].

[7] D. Gillet, "Personal learning environments: learners in control", In "Learning with Technology", 2010, pp. 193-201.

[8] M. Grahame Moore and O. Marty, https://www.researchgate.net/publication/281383979_La_theorie_de_la_distance_transactionnelle. Ed. researchgate.net. [retrieved: January, 2024].

[9] A. Tricot, "Educational innovation: myths and realities", Ed. Retz, 2017, p.160.

[10] https://www.canotech.fr/a/enseigner-avec-le-numerique-esprit-critique-et-travail-collaboratif. Ed. Canotech, Canopé. [retrieved: May, 2023].

[11] A. Mucchielli and C. Noy, "Communication studies: Constructivist approaches", Collection U, Ed. Armand Colin, 2005, p.240, chap. 9, pp. 165-189.

[12] E Morin, The seven pieces of knowledge known for future education, Ed. Seuil- UNESCO 1999, p.71.

[13] F. Singly (De), "The survey and its methods: the questionnaire", Ed. Nathan University, 1992, pp. 41-43.

[14] H. Hammadi, "Digital uses and practices in secondary school Achieving digital literacy - Digital collaboratif space (DCS) ", 2023. Ed.https://www.thinkmind.org/index.php?view=article&articleid=iciw_2023_1_20_20016. [retrieved: March, 2024].

 ISBN: 978-1-68558-147-3

On the Distribution of the Queue Size in a Packet Buffer

Blazej Adamczyk and Andrzej Chydzinski
Department of Computer Networks and Systems
Silesian University of Technology
Gliwice, Poland
email: {blazej.adamczyk, andrzej.chydzinski}@polsl.pl

***Abstract*—In this paper, we address the performance of the packet queueing mechanism at the output interface of a router. Specifically, we derive transient and stationary distribution of the queue size, which are fundamental performance characteristics of every queueing mechanism. The results are obtained directly in the time domain, without previous application of the Laplace transform, which is usually the case in the transient analysis. The model can be used for both passive buffer management (tail drop) and active buffer management, in which the dropping probability is a function of the queue size. It can be used also for traffic types of different statistical properties. Theoretical results are illustrated with numerical examples.**

Keywords-packet buffer; queue size distribution; passive buffer management; active buffer management.

I. INTRODUCTION

The buffers for packets, found in networking equipment, serve the purpose of temporarily storing sudden surges of packets caused by unpredictable peaks of network traffic. The buffers are integral components of packet networks. In their absence, the utilization of physical links between nodes would experience a notable decrease.

The size of the queue in a buffer is an important performance parameter. This is due to the fact that the time a packet spends in the buffer extends its total trip time to destination. Hence, several mathematical models have been proposed for calculations of the queue size, see, e.g., the classic monograph [1]. The classic models however, like the well-known M/M/1/N or M/G/1/N, constitute rather rough approximation of reality. They do not reflect important properties of real traffic and/or the packet service process at the router. For instance, they allow neither for modelling of an arbitrary packet interarrival time distribution nor its autocorrelated structure. Moreover, the classic analysis was focused on stationary characteristics mostly, without the full transient solution of the model.

Therefore, a few more complex queueing models were proposed and solved, with focus on advanced traffic modelling, e.g., [2]–[4], and the transient analysis, e.g., [5]–[7].

Recently, the situation got more complicated when active buffer management was advocated by the Internet Engineering Task Force (IETF) [8][9]. Most packet buffers exploited nowadays are governed by the passive, tail-drop algorithm [10] [11], i.e. incoming packets are dropped when the buffer is full. In active buffer management, the packets are dropped before the buffer gets full, with probability evolving in time. Many such schemes have been proposed for router's buffers, see, e.g., [12]–[19]. The most straightforward of these algorithms are based on the concept that the dropping probability should be a function of the queue size, see, e.g., [20]–[25].

This created a new research problem, i.e. finding the queue size distribution in a buffer with active management. The problem has been already solved to a large extent for algorithms exploiting the queue-size based dropping, namely, the solutions for simple traffic models were given in, e.g., [26]–[32], while for advanced traffic models in [33]–[38]. Moreover, works [28][29][32]–[38] encompass the transient analysis of various characteristics.

Now, in the transient analysis, the typical approach is based on moving equations to the Laplace transform domain [29] [33][38], solving them in this domain, and returning to the time domain with the help of the transform inversion algorithm. This method was successful in solving many queueing problems, but it can be viewed as rather complicated.

In this paper, we derive the transient queue size distribution directly in the time domain, without the help of the Laplace transform. Moreover, we give the formulas for the stationary distribution, and for the average queue size, both in the transient and the stationary case.

The considered model encompasses the active buffer management, in which the dropping probability is a function of the queue size. However, by a proper parameterization, it can be also used for calculations in the passive management case. The packet arrival stream is modeled by the Markovian arrival process [39], which is very robust as it can approximate with high accuracy any interarrival time distribution, with any autocorrelation. The function for packet dropping probabilities is general in the model and can assume any form.

The only cost we have to pay for these direct derivations is that the service time distribution (which is proportional to the packet size distribution) is approximated by the exponential distribution. Fortunately, this should not constitute a big problem in practice for the following reason. Packet sizes are strictly limited by the maximum transmission unit (MTU), e.g., in the 40-1500 bytes range. Therefore, if only the traffic is not dominated by small packets, then the coefficient of variation of the packet size is less than 1. Consider, for instance, the following pessimistic scenario: 50% of packets are of size 40 bytes, 50% of size 1500 bytes. The average packet size is in this case 770 bytes, while the coefficient of variation is 0.948. The analogous coefficient of variation for the exponential distribution is 1. Therefore, we can expect the queue sizes obtained using the exponential service time distribution, to be not far from the real ones, in this pessimistic

 ISBN: 978-1-68558-147-3

scenario. In reality, the coefficient of variation of the packet size is usually significantly less than 1, often less than 0.5. Therefore, the discrepancy between the queue sizes obtained using the exponential approximation should be on the side of pessimistic overestimation.

The rest of the paper is structured as follows. In Section II, the details of the queueing model are presented. In Section III, the actual analysis of the queue size distribution, its average value and the standard deviation, are carried out. The section is divided into two parts, devoted to the steady-state analysis, and the transient analysis, respectively. In Section IV, numerical examples are presented. They include stationary and transient results, with full distributions accompanied by average values and standard deviations. Moreover, the convergence to the stationary state is demonstrated. The final conclusions are gathered in Section V.

II. The model

The buffer is modeled herein by the single-server queueing system of finite capacity, namely, the packets arrive to the buffer according to the arrival process, which will be defined below. In the buffer, they form a queue in the arrival order, in a First-In-First-Out (FIFO) manner. The packets are served and removed from the head of the queue by the transmission process. Packet transmission time is random and exponentially distributed with parameter μ. The capacity of the system (buffer) is K packets, which includes the one being transmitted, if applicable. If upon a packet arrival the buffer is full, the arriving packet is dropped. Moreover, every arriving packet can be dropped even if the buffer is not full. This happens with probability $d(n)$, where n is the queue size upon arrival of this packet. Function $d(n)$ can have any form if only it meets $0 \leq d(n) \leq 1$ for every n.

The packet arrival proces is modeled by the Markovian arrival process [39]. This process has an internal modulating Markov chain with m states, which can modulate the actual interarrival times to have a complicated form of the distribution of interarrival times and autocorrelation.

In practice, the Markovian arrival process is parameterized by two $m \times m$ matrices, D_0 and D_1. Diagonal elements of D_1 cover arrivals of packets without switching the modulating state, while off-diagonal elements cover arrivals of packets accompanied by switching the modulating state. Off-diagonal entries of D_0 cover switching the modulating state without arrivals.

More on the properties and detailed characteristics of the Markovian arrival process can be found in [39].

In what follows, by $X(t)$ we will denote the queue size (in packets) at the time t, including the one being transmitted, if applicable. By $J(t)$ we will denote the modulating state at the time t, i.e. the state of the internal Markov chain. The space of possible states is $\{1, \ldots, m\}$.

III. Queue size analysis

First of all, we can notice that the two-dimensional process, $(X(t), J(t))$ constitutes a continuous-time Markov chain in the considered model.

Indeed, at any particular moment in time, t, the evolution of the arrival process after t depends only on $J(t)$, i.e. the modulating state at t. It does not depend on the values of the modulating state before t.

If there is an ongoing service at t, the remaining service time is exponentially distributed with parameter μ, which is a consequence of the memoryless property of the exponential distribution. In other words, the distribution of the remaining service time counting from t does not depend on already passed service.

Finally, future dropping of packets, counting from t, depends only on the current queue size, $X(t)$.

Summarizing, the evolution of the system counting from t depends only on the current queue size, $X(t)$, and the current modulating state, $J(t)$, which makes the process $(X(t), J(t))$ to be a continuous-time Markov chain.

Let Q be the rate matrix of this two-dimensional Markov chain. Obviously, Q must be of size $(K+1)m \times (K+1)m$, because it covers simultaneously changes of the queue size, with possible values $0, \ldots, K$ and changes of the modulating state, with possible values $1, \ldots, m$.

We will derive now the matrix Q.

Firstly, note that when the system is empty, the change of the modulating state without changing the queue size can happen either when there is no packet arrival, which is covered by D_0, or when an arriving packet gets dropped immediately, which is covered by $d(0)D_1$. Therefore, the two possibilities together are covered by the matrix:

$$D_0 + d(0)D_1. \tag{1}$$

When the queue size is $i > 0$, the change of the modulating state without changing the queue size can happen either when there is no packet arrival, which is covered by D_0, or when a packet is dropped instantly, which is covered by $d(i)D_1$. Moreover, we have to exclude the service completion events, happening with intensity μ. Therefore, the three possibilities together are covered by the matrix:

$$D_0 + d(i)D_1 - \mu I, \tag{2}$$

where I is the $m \times m$ identity matrix.

A successful packet arrival event, with perhaps a change of the modulating state, is covered by the matrix:

$$(1 - d(i))D_1. \tag{3}$$

In this case, the queue size increases by 1.

Then, any non-zero queue size can be decreased by 1 at any time by the service process. This can happen only without a change of the modulating state and is covered by the matrix:

$$\mu I. \tag{4}$$

In the considered model, instant changes of the queue size by more than 1 are impossible, neither up nor down. Therefore, such events have probability 0.

Summarizing these considerations, we obtain matrix Q in the following form:

$$Q = [Q_{ij}]_{i,j=0,\ldots,K}, \tag{5}$$

 ISBN: 978-1-68558-147-3

$$Q_{ij} = \begin{cases} D_0 + d(0)D_1, & \text{if } i = 0, j = 0, \\ D_0 + d(i)D_1 - \mu I, & \text{if } i = j, i > 0, \\ (1 - d(i))D_1, & \text{if } i = j - 1, \\ \mu I, & \text{if } i = j + 1, \\ \mathrm{O}, & \text{if } i > j + 1, \\ \mathrm{O}, & \text{if } i < j - 1, \end{cases} \tag{6}$$

where each entry of Q is an $m \times m$ submatrix, giving its total size $(K+1)m \times (K+1)m$. O in (6) denotes the zero matrix.

It is also useful to present Q in a graphical form. Namely, from (6) we have:

$$Q = \begin{pmatrix} U & Z_0 & \mathrm{O} & \mathrm{O} & \cdots & \mathrm{O} \\ \mu I & Y_1 & Z_1 & \mathrm{O} & \cdots & \mathrm{O} \\ \mathrm{O} & \mu I & Y_2 & Z_3 & \cdots & \mathrm{O} \\ \mathrm{O} & \mathrm{O} & \mu I & Y_3 & \ddots & \mathrm{O} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathrm{O} & \mathrm{O} & \mathrm{O} & \cdots & \mu I & Y_K \end{pmatrix}, \tag{7}$$

with

$$U = D_0 + d(0)D_1, \tag{8}$$

$$Y_i = D_0 + d(i)D_1 - \mu I, \quad i \geq 1, \tag{9}$$

$$Z_i = (1 - d(i))D_1, \quad i \geq 0. \tag{10}$$

A. Stationary solution

Having rate matrix Q, we can obtain the stationary distribution of the queue size and the modulating state:

$$q_{nj} = \lim_{t \to \infty} \mathbb{P}(X(t) = n, J(t) = j | X(0) = k, J(0) = i), \tag{11}$$

using the system of linear equations:

$$qQ = [0, \ldots, 0], \quad \sum_{n=0}^{K} \sum_{j=1}^{m} q_{nj} = 1, \tag{12}$$

where

$$q = [q_{01}, \ldots, q_{0m}, q_{11}, \ldots, q_{1m}, \ldots, q_{K1}, \ldots, q_{Km}]. \tag{13}$$

It is known that system (13) has a unique solution, if the Markov chain is finite and aperiodic, as in our case [1]. It is also known that this solution does not depend on initial conditions, $X(0) = k$ and $J(0) = i$.

Having computed vector q, we can obtain easily the distribution of the queue size in the stationary state. Defining:

$$p_n = \lim_{t \to \infty} \mathbb{P}(X(t) = n | X(0) = k, J(t) = i), \tag{14}$$

we have

$$p_n = \sum_{j=1}^{m} q_{nj}. \tag{15}$$

The average queue size in the stationary state, A, equals:

$$A = \sum_{n=0}^{K} n \sum_{j=1}^{m} q_{nj}, \tag{16}$$

while the standard deviation of the queue size in the stationary state, S, is:

$$S = \sqrt{\sum_{n=0}^{K} n^2 \sum_{j=1}^{m} q_{nj} - A^2}. \tag{17}$$

Finally, the stationary state distribution of the modulating state can be also obtained. Namely, we have:

$$r_j = \lim_{t \to \infty} \mathbb{P}(J(t) = j | X(0) = k, J(t) = i) = \sum_{n=0}^{K} q_{nj}. \tag{18}$$

B. Transient solution

Having rate matrix Q, we can obtain also the distribution of the queue size and the modulating state at any time. Defining:

$$q_{kinj}(t) = \mathbb{P}(X(t) = n, J(t) = j | X(0) = k, J(t) = i), \tag{19}$$

we have:

$$q_{kinj}(t) = \left[e^{Qt}\right]_{(k \cdot m + i, n \cdot m + j)}, \tag{20}$$

where e^{Qt} is the matrix exponential, while

$$\left[M\right]_{(a,b)}$$

denotes the (a, b) entry of matrix M.

From $q_{kinj}(t)$, we can obtain the distribution of the queue size at the time t, i.e.:

$$p_{kin}(t) = \mathbb{P}(X(t) = n | X(0) = k, J(t) = i). \tag{21}$$

Namely, we have

$$p_{kin}(t) = \sum_{j=1}^{m} q_{kinj}(t). \tag{22}$$

Finally, defining $A_{ki}(t)$ to be the average queue size at the time t, given that $X(0) = k$ and $J(t) = i$, we obtain:

$$A_{ki}(t) = \sum_{n=0}^{K} n \sum_{j=1}^{m} q_{kinj}(t). \tag{23}$$

The standard deviation of the average queue size at the time t, given that $X(0) = k$ and $J(t) = i$, is equal to:

$$S_{ki}(t) = \sqrt{\sum_{n=0}^{K} n^2 \sum_{j=1}^{m} q_{kinj}(t) - (A_{ki}(t))^2}. \tag{24}$$

C. Special case - passive buffer management

It is easy to see that the presented model can be applied to the passive buffer management as well. If we set:

$$d(n) = \begin{cases} 0, & \text{for} \quad n < K, \\ 1, & \text{for} \quad n \geq K, \end{cases} \tag{25}$$

than the resulting model is equivalent to the classic, tail-drop buffer management. All the presented derivations and formulas remain valid, because in the definition of the model given in Section II, function $d(n)$ has an arbitrary form.

D. Special cases - arrival processes of various types

The complex Markovian arrival process assumed in the definition of the model in Section II can be simplified when needed. For instance:

- setting $m = 1$ and $D_0 = -\lambda$, $D_1 = \lambda$, we get the Poisson process;
- setting $D_0 = T$ and $D_1 = -T\mathbf{1}\alpha$, we get the renewal process, in which the interarrival time distribution is of phase type (a phase-type distribution can approximate any distribution with arbitrary accuracy);
- setting $D_1 = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ we get the Markov-modulated Poisson process, which is a simple and popular model of autocorrelated traffic.

Again, all the presented derivations and formulas remain valid in the each case listed above.

IV. Numerical Examples

In the examples, the following parameterization of the Markovian arrival process is used:

$$D_0 = \begin{bmatrix} -0.4395723602 & 0.03517495366 & 0.01134675916 \\ 0.04652171271 & -0.6814139155 & 0.04652171271 \\ 0.01248143512 & 0.01248143523 & -2.4156995522 \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0.21503476203 & 0.09914223374 & 0.07887365167 \\ 0.04725831044 & 0.48250711000 & 0.05860506971 \\ 0.13318251133 & 0.04765638867 & 2.20989778185 \end{bmatrix}.$$

These matrices describe a moderately autocorrelated stream, with the 1-lag autocorrelation of 0.188 and the rate of 1.1. Moreover, it is assumed that the packet transmission rate is 1. Thus, the queue is slightly overloaded, with $\rho = 1.1$. To reduce this overload, the following active buffer management is used in the system (see Figure 1):

$$d(n) = \begin{cases} 0, & \text{for } n < 5, \\ 0.0002(n-5)^3, & \text{for } 5 \le n < 20, \\ 1, & \text{for } n \ge 20, \end{cases} \quad (26)$$

which is a third-degree polynomial, suggested in [22]. The buffer size is $K = 20$.

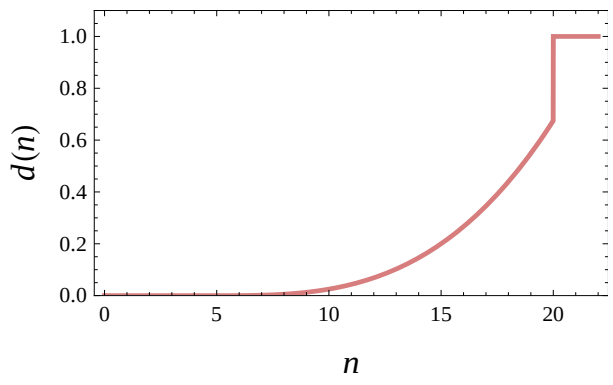

Figure 1. Function $d(n)$ used in numerical examples.

In Table I, the average queue size and its standard deviation are presented at different moments in time. (The initial queue size was 0, while the initial modulating state was 1 in this calculations). The results for t up to 1000 were obtained from formulas (23) and (24), while the results for $t = \infty$ were obtained from (16) and (17), respectively.

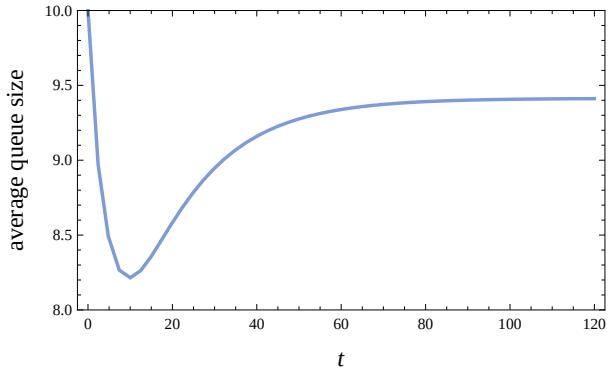

Figure 2. Average queue size in time for $X(0) = 10$ and $J(0) = 1$.

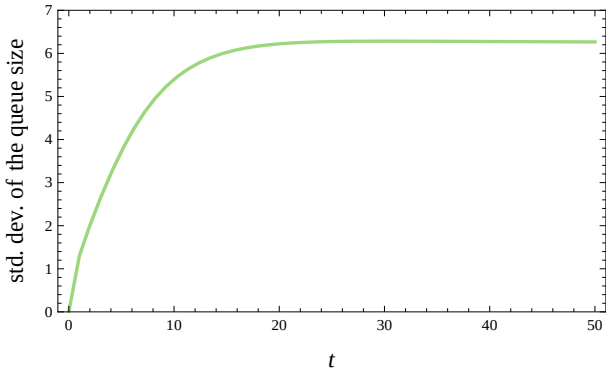

Figure 3. Std. dev. of the queue size in time for $X(0) = 10$ and $J(0) = 1$.

TABLE I
AVERAGE QUEUE SIZE AND ITS STANDARD DEVIATION IN TIME FOR $X(0) = 0$ AND $J(0) = 1$.

	average queue size at t	std. dev. of the queue size at t
$t = 0.1$	0.038	0.201
$t = 1.0$	0.346	0.734
$t = 10$	3.630	4.253
$t = 100$	9.389	6.254
$t = 1000$	9.413	6.251
$t = \infty$	9.413	6.251

In Figures 2 and 3, we can see the evolution of the average queue size and its standard deviation in time for the initial queue size of 10 and the initial modulating state of 1.

As we can see in Table I and Figure 2, the full convergence of the average value to the stationary state takes about 120s, while the standard deviation stabilized much quicker than that. It is interesting that in the case of the initial queue of 10 packets (i.e. 50% of the buffer), the average queue size is not monotonic (see Figure 2). It is decreasing during the first 10s interval, then begins to increase until the stationary state is

 ISBN: 978-1-68558-147-3

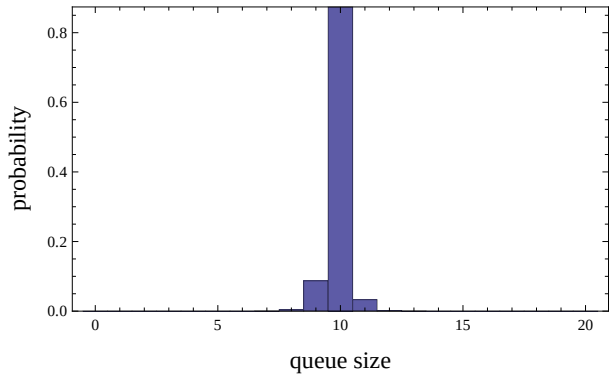

Figure 4. Distribution of the queue size at $t = 0.1$. Initial conditions: $X(0) = 10$ and $J(0) = 1$.

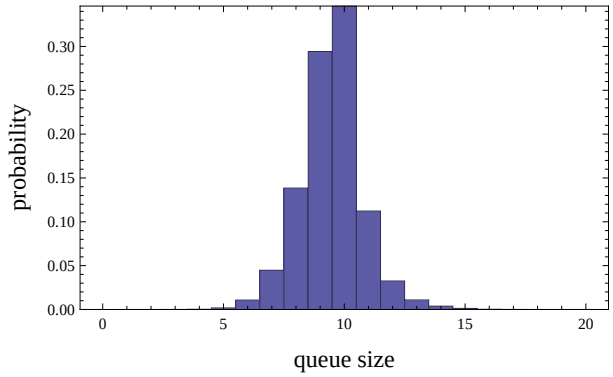

Figure 5. Distribution of the queue size at $t = 1.0$. Initial conditions: $X(0) = 10$ and $J(0) = 1$.

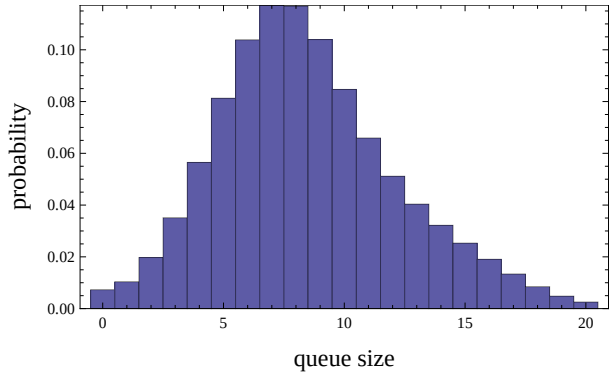

Figure 6. Distribution of the queue size at $t = 5.0$. Initial conditions: $X(0) = 10$ and $J(0) = 1$.

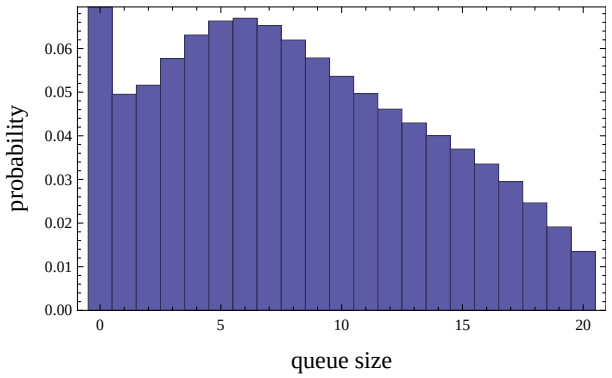

Figure 7. Distribution of the queue size at $t = 10$. Initial conditions: $X(0) = 10$ and $J(0) = 1$.

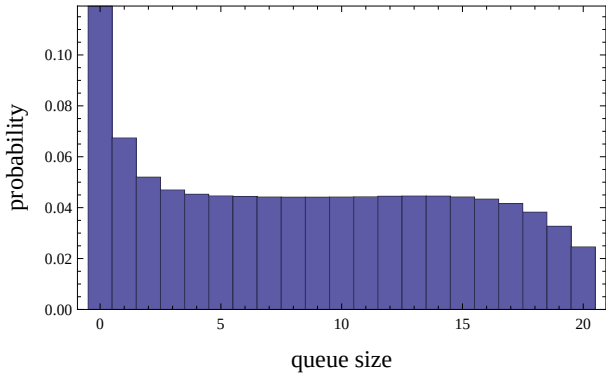

Figure 8. Distribution of the queue size at $t = 20$. Initial conditions: $X(0) = 10$ and $J(0) = 1$.

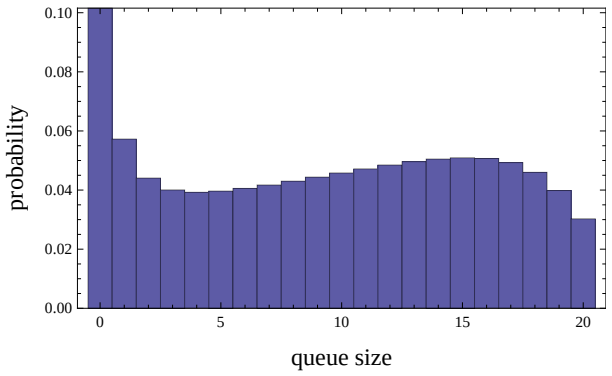

Figure 9. Stationary distribution of the queue size ($t = \infty$).

reached.

Now, in Figures 4–8, the full distribution of the queue size is depicted at different times, while in Figure 9 the stationary distribution is shown. Figures 4–8 were obtained by means of formulas (22) and (20), while Figure 9 by means of (12) and (15).

In these figures, we can track the evolution of the shape of the queue size distribution towards the stationary distribution. Namely, very early, the probability mass is concentrated around the initial queue size, equal 10 (see Figures 4 and 5). Then, it becomes more spread out – see Figure 6. After some more time, a peak at 0 occurs (Figure 7). At $t = 20$, the distribution starts to resemble quite well the stationary distribution – compare Figures 8 and 9.

 ISBN: 978-1-68558-147-3

V. CONCLUSIONS

In this paper, we derived transient and stationary distribution of the queue size in a packet buffer, as well as its average value and standard deviation. The derivations were carried out in the time domain, without previous application of the Laplace transform. The analyzed model can be used for both passive buffer management (tail drop) and active buffer management, in which the dropping probability is a function of the queue size. It can be used for many traffic types, of different statistical properties, including the Poisson process, phase-type renewal process, Markov-modulated Poisson process, and others. Theoretical results are illustrated with numerical examples.

The only significant simplification of the model was approximation of the service time by the exponential distribution. This should not constitute a big problem in practice, because the real coefficient of variation of the packet size is usually smaller than the coefficient of variation of the exponential distribution. Therefore, the error caused by this approximation should be on the side of pessimistic overestimation of the queue size.

ACKNOWLEDGEMENT

This research was funded by National Science Centre, Poland, grant number 2020/39/B/ST6/00224.

REFERENCES

[1] L. Kleinrock, "Queueing Systems: Theory", John Wiley and Sons New York, 1975.

[2] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process", Commun. Stat., Stochastic Models, vol. 7, no. 1, pp. 1–46, 1991.

[3] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook", Performance Evaluation, vol. 18, no. 2, pp. 149–171, 1992.

[4] A. N. Dudin, A. A. Shaban, and V. I. Klimenok, "Analysis of a queue in the BMAP/G/1/N system", International Journal of Simulation: Systems, Science and Technology, vol. 6, no. 1–2, pp. 13–23, 2005.

[5] T. Hofkens, K. Spaey, and C. Blondia, "Transient Analysis of the D-BMAP/G/1 Queue with an Application to the Dimensioning of a Playout Buffer for VBR Video", Proceedings of Networking'04. Lecture Notes in Computer Science, vol. 3042, pp. 1338–1343, 2004.

[6] A. Chydzinski, "Transient Analysis of the MMPP/G/1/K Queue", Telecommunication Systems, vol. 32, no. 4, pp. 247–262, 2006.

[7] A. Chydzinski, "On the Queue Length Distribution in BMAP Systems", Journal of Communications Software and Systems, vol. 2, no. 4, pp. 275–282, 2006.

[8] Internet Engineering Task Force, "Request for Comments 7567: ETF Recommendations Regarding Active Queue Management" F. Baker and G. Fairhurst, Eds.; 2015.

[9] Internet Engineering Task Force, "Request for Comments 7928: Characterization Guidelines for Active Queue Management (AQM)", N. Kuhn, F. Natarajan, and N. Khademi, Eds.; 2016.

[10] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers", ACM SIGCOMM Computer Communication Review, vol. 34, pp. 281–292, 2004.

[11] B. Spang, S. Arslan, and N. McKeown, "Updating the theory of buffer sizing", Performance Evaluation, vol. 151,102232, 2021.

[12] K. Nichols and Van Jacobson, "Controlling Queue Delay", Queue, vol. 55, no. 7, pp. 42–50, 2012.

[13] R. Pan, P. Natarajan, C. Piglione, and M. S. Prabhu, "PIE: A lightweight control scheme to address the bufferbloat problem", IEEE International Conference on High Performance Switching and Routing, pp. 148–155, 2013.

[14] G. Kahe and A. H. Jahangir, "A self-tuning controller for queuing delay regulation in TCP/AQM networks", Telecommunication Systems, vol. 71, no. 2, pp. 215–229, 2019.

[15] S. B. Alaoui, E. H. Tissir, and N. Chaibi, "Analysis and design of robust guaranteed cost Active Queue Management", Computer Communications, vol. 159, pp. 124–1321, 2020.

[16] H. Wang, "Trade-off queuing delay and link utilization for solving bufferbloat", ICT Express, vol. 6, no. 4, pp. 269–272, 2020.

[17] R. Hotchi, "Active queue management supporting TCP flows using disturbance observer and Smith predictor", IEEE Access, vol. 8, pp. 173401–173413, 2020.

[18] C. Wang, X. Chen, J. Cao, and J. Qiu, "Neural Network-Based Distributed Adaptive Pre-Assigned Finite-Time Consensus of Multiple TCP/AQM Networks", IEEE Transactions on Circuits and Systems, vol. 68, no. 1, pp. 387–395, 2021.

[19] J. Shen, Y. Jing, and T. Ren, "Adaptive finite time congestion tracking control for TCP/AQM system with input saturation", International Journal of Systems Science, vol. 53, no. 2, pp. 253–264, 2022.

[20] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", IEEE/ACM Transactions on Networking, vol. 1, no. 4, pp. 397–413, 1993.

[21] K. Zhou, K. L. Yeung, and V. Li, "Nonlinear RED: asimple yet efficient active queue management scheme", Computer Networks, vol. 50, pp. 3784–3784, 2006.

[22] J. Domanska, D. Augustyn, and A. Domanski, "The choice of optimal 3-rd order polynomial packet dropping function for NLRED in the presence of self-similar traffic", Bulletin of the Polish Academy of Sciences. Technical Sciences, vol. 60, no. 4, pp. 779–786, 2012.

[23] A. Gimenez, M. A. Murcia, J. M. Amigo, O. Martinez-Bonastre, and J. Valero, "New RED-Type TCP-AQM Algorithms Based on Beta Distribution Drop Functions", Applied Sciences, vol. 12, no. 21, 2022.

[24] C. Feng, L. Huang, C. Xu, and Y. Chang, "Congestion Control Scheme Performance Analysis Based on Nonlinear RED" IEEE Systems Journal, vol. 11, no. 4, pp. 2247–2254, 2017.

[25] S. Patel and Karmeshu, "A New Modified Dropping Function for Congested AQM Networks", Wireless Personal Communications, vol. 104, no. 1, pp. 37–55, 2019.

[26] T. Bonald, M. May, and J.-C. Bolot, "Analytic evaluation of RED performance", Proceedings of INFOCOM, pp. 1415–1424, 2000.

[27] A. Chydzinski and L. Chrost, "Analysis of AQM queues with queue-size based packet dropping", International Journal of Applied Mathematics and Computer Science, vol. 21, no. 3, pp. 567-–577, 2011.

[28] W. M. Kempa, "A direct approach to transient queue-size distribution in a finite-buffer queue with AQM", Applied Mathematics & Information Sciences, vol. 7, no. 3, pp. 909-915, 2013.

[29] W. M. Kempa, "Time-dependent queue-size distribution in the finite GI/M/1 model with AQM-type dropping", Acta Electrotech. Et Inform., vol. 13, pp. 85–90. 2013.

[30] A. Chydzinski, M. Barczyk, and D. Samociuk, "The Single-Server Queue with the Dropping Function and Infinite Buffer", Mathematical Problems in Engineering, vol. 2018, ID 3260428, pp. 1–12, 2018.

[31] A. Chydzinski, "Queues with the dropping function and non-Poisson arrivals", IEEE Access, vol. 8, no. 1, pp. 39819–39829, 2020.

[32] A. Chydzinski, "Non-Stationary Characteristics of AQM Based on the Queue Length", Sensors, vol. 23, no. 1, 485, 2023.

[33] P. Mrozowski and A. Chydzinski, "Queues with dropping functions and autocorrelated arrivals", Methodology and Computing in Applied Probability, vol. 20, no. 1, pp. 97–115, 2018.

[34] A. Chydzinski and B. Adamczyk, "Queues with the dropping function and general service time", PLoS ONE, vol. 14, no. 7, e0219444, pp.1-21, 2019.

[35] A. Chydzinski, "Waiting Time in a General Active Queue Management Scheme", IEEE Access, vol. 11, pp. 66535-66543, 2023.

[36] A. Chydzinski, "Output Stream from the AQM Queue with BMAP Arrivals", Journal of Sensor and Actuator Networks, vol. 13(1), no. 4, pp. 1-18, 2024.

[37] A. Chydzinski, "Throughput of the Queue With Probabilistic Rejections", IEEE Access, vol. 11, pp. 138141-138150, 2023.

[38] A. Chydzinski, "Loss Process at an AQM Buffer", Journal of Sensor and Actuator Networks, vol. 12(4), no. 55, pp. 1-17, 2023.

[39] A. N. Dudin, V. I. Klimenok, and V. M. Vishnevsky, "The Theory of Queuing Systems with Correlated Flows", Springer Cham, 2020.

 ISBN: 978-1-68558-147-3

A Short Survey on Graph Neural Networks Based Stock Market Prediction Models

Lei Zhou
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
lei.zhou@autuni.ac.nz

Yuqi Zhang
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
yuqi.zhang@autuni.ac.nz

Jian Yu
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
jian.yu@aut.ac.nz

Sira Yongchareon
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
sira.yongchareon@aut.ac.nz

Samaneh Madanian
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
sam.madanian@aut.ac.nz

Guiling Wang
School of Information Science and Technology
North China University of Technology
Beijing, China
wangguiling@ncut.edu.cn

***Abstract*—Stock market predictions are challenging due to the complexity and volatility of markets across the globe. Influential factors include, but are not limited to, economic conditions, political events, investor sentiment, and even natural calamities. In this survey, we review the current literature on stock market predictions using various approaches and propose a framework that facilitates the categorization and analysis of existing works. A novel taxonomy is also proposed within this framework for Graph Neural Network (GNN)-based stock market prediction methods. Potential research gaps are identified, and future research directions are discussed towards the end of this survey.**

***Index Terms*—Graph neural networks, stock prediction, deep learning.**

I. Introduction

The stock market's complexity and dynamic nature necessitate accurate, interpretable prediction models. Traditional forecasts using time series models such as ARIMA [1] and GARCH [2] face limitations due to the market's nonlinear evolution.

Deep learning, applicable across fields including Natural Language Processing(NLP) and financial forecasting, surpasses traditional methods by handling diverse data types and capturing non-linear stock relationships. Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory(LSTM) [3] and Gated Recurrent Unit(GRU) [4], excel in time series but struggle with inter-stock dynamics. Graph Convolutional Networks (GCN) [5] enable relational reasoning, improving performance by incorporating stock relationships.

To use Graph Neural Networks (GNN) [6] for stock prediction, one can represent stocks as nodes to facilitate node classification tasks. This leverages the stock relationships to enhance prediction models. Chen et al. [7] demonstrated GNN's potential in stock market representation and prediction.

Despite extensive surveys on machine learning and deep learning in stock prediction [8]–[15], a gap exists in dedicated GNN applications. This survey addresses graph-based deep learning advancements in stock prediction, proposing a GNN classification framework, a novel taxonomy, identifying research gaps, and suggesting further directions.

This paper is organized as follows: Section II introduces prediction methods and our framework. Section III reviews GNN applications and our taxonomy. Section IV discusses prior research, and Section V concludes.

II. Classification Framework

In this section, we propose a novel classification framework that analyses existing approaches from three aspects, including *model architecture, dataset feature and graph construction*. Figure 1 illustrates the proposed taxonomy based on the classification framework.

A. Model Architecture.

We have identified three types of model architectures that can represent most existing literature: (1) RNN-GNN architecture, (2) iterative RNN-GNN architecture, and (3) parallel RNN-GNN architecture. Figure 2 shows the structure of the three architectures. The RNN-GNN architecture is the most commonly used for graph-based stock market prediction. Stock time series data are first fed into a Recurrent Neural

 ISBN: 978-1-68558-147-3

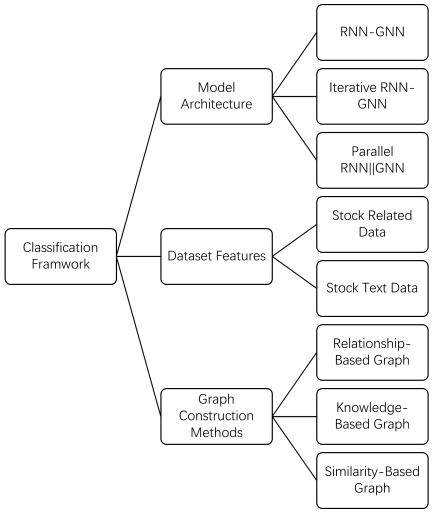

Fig. 1: Classification Framework.

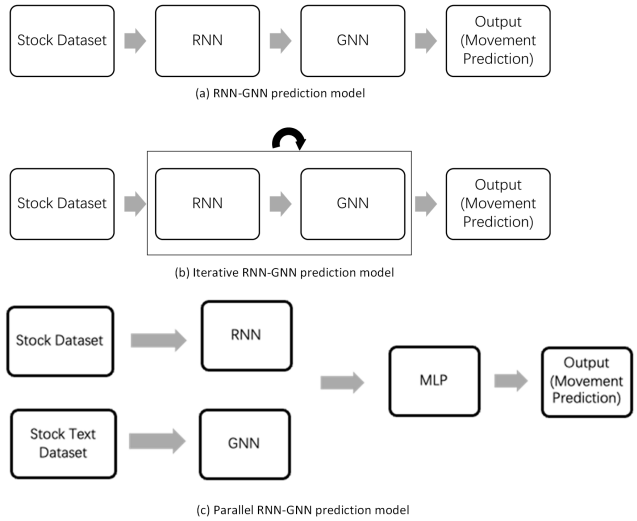

Fig. 2: Common architecture using GNN for stock prediction.

Network, such as LSTM [3], GRU [4], or Bi-directional [16], to extract features. Then, the hidden features are processed by a Graph Neural Network, such as GCN [5], Graph Attention Network(GAT) [17], or GraphSAGE [18], for node classification. The other two architectures are less common in the literature.

As shown in Figure 2 (b), the iterative RNN-GNN architecture achieves deeper integration and collaboration between the RNN and GNN, including an information exchange mechanism, resulting in better capture of the complex relationships between temporal and relational data.

The parallel RNN ‖ GNN architecture, shown in Figure 2 (c), effectively captures both temporal and relational information in complex datasets. This integration facilitates a more comprehensive understanding of the data, leading to improved performance in tasks such as sequence modeling, time series forecasting, and relational reasoning.

Compared with the other two architectures, the RNN-GNN model is simpler and easier to implement, as it uses the output of the RNN directly as input to the GNN. However, it may underperform in some complex tasks. The choice of architecture depends on task-specific requirements and data characteristics.

B. Datasets Features

Stock price fluctuations depend on numeric and text data, including prices and volumes [19], as well as announcements and social media [20]. Using historical prices helps predict future trends, with numeric data being more standardized and accessible. Text data, requiring sentiment analysis for integration [21], contributes to forecasting but cannot solely predict prices. Combining both data types enhances prediction accuracy.

C. Graph Construction Method

In literature on GNN for stock market prediction, three primary methods for graph construction are identified: correlation-based, knowledge-based, and similarity-based graph constructions.

Relationship-based graph construction [7] constructs graphs from internal stock relationships, utilizing key market indicators such as prices and volumes to create nodes for each stock. Historical data analysis, through correlation coefficients or time series models, determines edges reflecting stock correlations.

Knowledge-based graph construction [22] [23] leverages external information, such as sector dependencies or economic impacts, to enhance graph structure. This approach integrates expert insights or industry reports, capturing relationships beyond dataset information.

Similarity-based graph construction [24] builds graphs by identifying similarities among stocks, useful for implicit relationship mapping. Nodes represent stocks, with edges based on similarity scores (e.g., cosine similarity or Euclidean distance), highlighting clusters of similar stocks.

III. Approaches

In this section, we review existing approaches that use GNNs for the stock market. We discuss each approach from three aspects: architecture, dataset features, and graph construction methods, as proposed in the previous section.

A. Model Architecture

1. RNN-GNN architecture

Most models adopt the RNN-GNN architecture, embedding time series data for graph-based prediction. Combining RNNs' ability to capture temporal sequences with GNNs' insight into stock interrelations, data is encoded using LSTM or GRU for pattern recognition and long-term dependencies.

Once the time series data is embedded, it is fed into the GNN component. The GNN utilizes graph convolutional layers

 ISBN: 978-1-68558-147-3

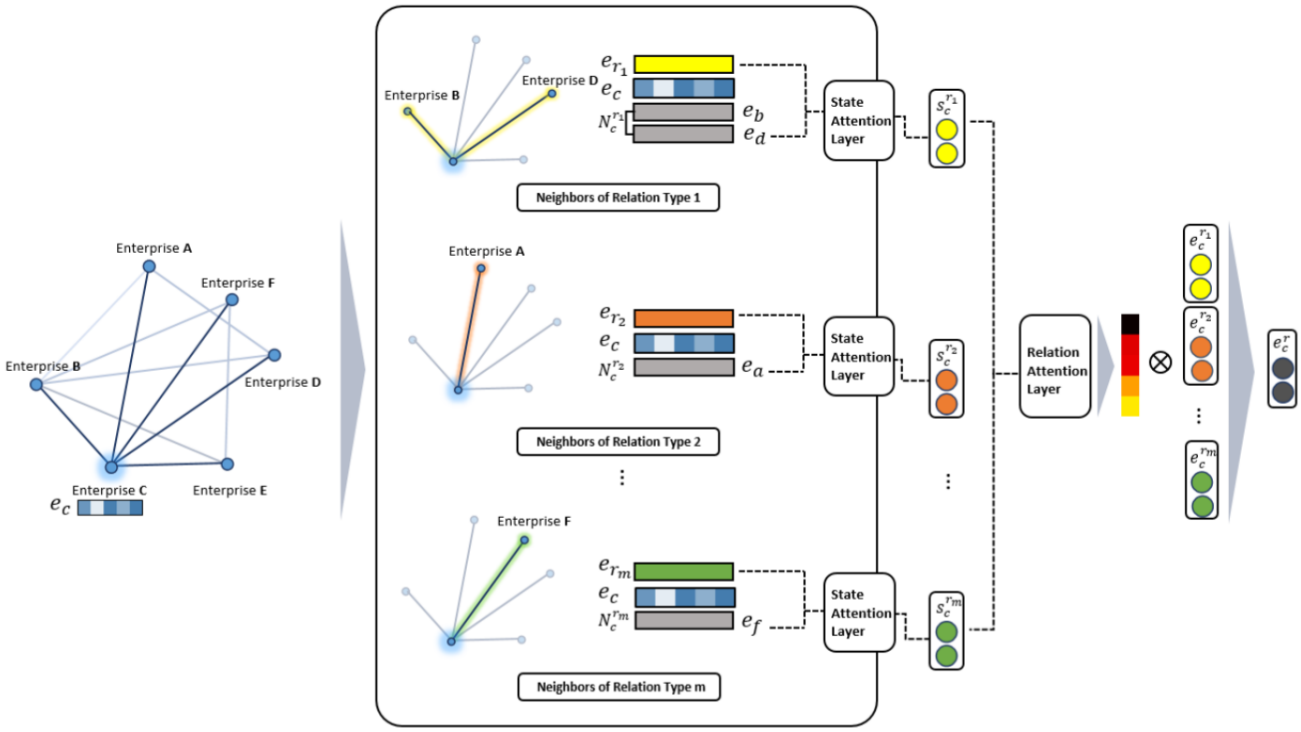

Fig. 3: Hierarchical Attention Network for Stock Prediction.

to propagate information among the interconnected stocks in a graph structure. By considering the correlations and dependencies between stocks, the GNN can capture collective behaviour and market dynamics.

In [7], Chen et al. propose a joint RNN-GNN model called the Incorporating Corporation Relationship-Graph Convolutional Neural Network (ICR-GCN), which employs the RNN-GNN architecture. The model is composed of two parts. The first part is an LSTM that encodes time series information to extract features for each company. These features then serve as the node attributes in a graph that represents the relationships between companies. Subsequently, a three-layer GCN is applied for node classification.

$$Y = \text{softmax}\left(\widehat{A}ReLU\left(\widehat{A}ReLU\left(\widehat{A}X'W\right)W\right)W\right)$$

where $\widehat{A}$ is the adjacency matrix, X' represents the historical features, W is the learnable weight matrix. $ReLU(\cdot)$ and $softmax$ functions are used as the activation functions, and cross-entropy is used as the loss function.

Compared to similar studies, a significant advancement in [25] is its consideration of social media text's impact on stock prices. By integrating social media text with financial data and stock relationships, it introduces additional dimensions of signals for stock prediction.

In [26], Kim et al. propose a Hierarchical Attention Network for Stock prediction (HATS) that uses relational data for stock market prediction. It selectively aggregates information on different relation types and adds the information to the representations of each company. As shown in Figure 3. HATS is the RNN-GNN model; it has three layers including the Feature Extraction layer, the Relational Modeling layer and the Prediction Layer. In the feature extraction layer, one of LSTM and GRU is used to encode features.

The Relational Modeling layer is used to encode the graph structure. The HATS layout is shown in Figure 3. e_{r_m} is the relation type, e_n is the feature of node n, and $N_i^{r_m}$ is the set of neighboring nodes of i for relation type m.

The prediction layer focuses on individual stock and S&P500 Index movements, classifying stocks into three categories: up, down, and neutral, using a linear transformation for individual predictions. Mean pooling calculates the index's graph representation. HATS, similar to ICR-GCN, employs RNN-GNN architecture but differs by using GAT, whereas ICR-GCN utilizes GCN.

HATS selectively aggregates information from different types of relationships and adds the information to the representation of each company.

In Feng et al.'s study [27], the Relational Stock Ranking (RSR) framework employs the Temporal Graph Convolution (TGC) model for stock prediction, featuring a three-layered joint RNN-GNN model: a sequential embedding layer with LSTM for capturing stock sequences, a relational embedding layer using TGC for stock interconnections, and a ranking scores prediction layer for stock ranking. Different from other models, RSR-TGC uniquely captures temporal dynamics with

 ISBN: 978-1-68558-147-3

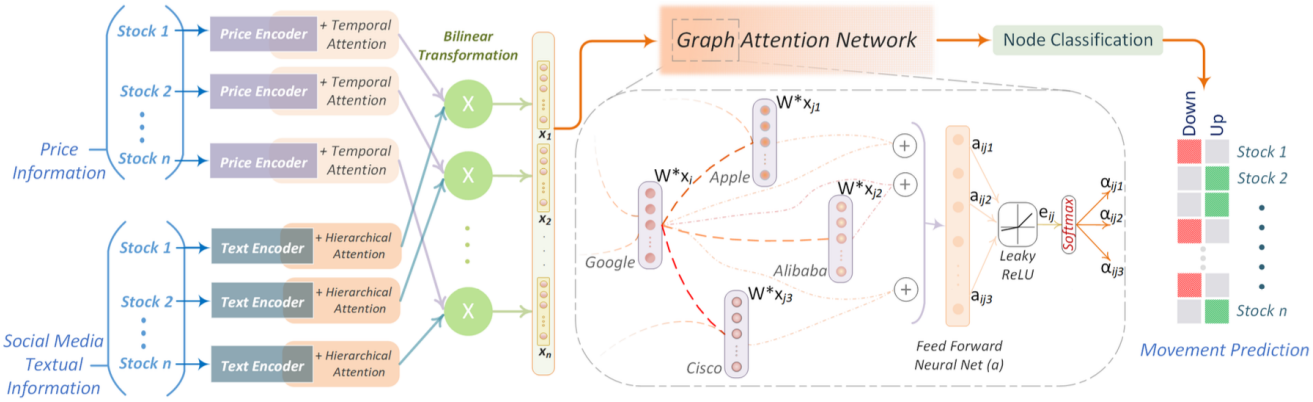

Fig. 4: MAN-SF Model.

TGC, distinguishing it from GCN and GAT models.

This approach contrasts with Wang et al. [28], which uses a hierarchical model to analyze temporal stock relationships and market dynamics across multiple timescales, considering financial, social media sentiment, and other factors.

In [25], Sawhney et al. propose a Multipronged Attention Network for Stock Forecasting (MAN-SF) by learning from historical prices, social media, and inter-stock relations. It is made up of a hierarchical attention network and a GAT. The Hierarchical Attention Network (HAN) is responsible for capturing relevant signals across diverse data, while the GAT is responsible for predicting stock movements.

MAN-SF model is a joint RNN-GNN architecture. As shown in Figure 4, first, GRU is used as a Price Encoder (PE) that takes the prices of a stock over a period of time and uses that to produce a price feature. The temporal attention mechanism is a way of aggregating information from different time steps into an overall representation. This is done by assigning learned weights to each time step, which allows the most important information to be aggregated together. For example, the formula of temporal attention mechanism $\zeta(\cdot)$ is shown as:

$$\zeta\left(\bar{h}_z\right) = \sum_i \beta_i h_i \tag{1}$$

$$\beta_i = \frac{\exp\left(h_i^T W \bar{h}_z\right)}{\sum_{i=1}^{T} \exp\left(h_i^T W \bar{h}_z\right)} \tag{2}$$

where $\bar{h}_z$ is the hidden states of GRU, β_i is the attention weight and W is the learnable parameter matrix.

Secondly, the Social Media Information Encoder (SMI) employs GRU to distill tweet data, using a hierarchical attention mechanism to encode this information into vectors. Thirdly, the Blending Multimodal Information layer merges features from PE and SMI, applying a bilinear transformation for learning price-tweet interactions, optimizing the mix of data inputs. Lastly, stock movement prediction is performed using a GAT.

Similar to the above models, MAN-SF also uses the RNN-GNN architecture pattern; the main difference is that three attention mechanisms are used to extract features, which include price data, news data and stock relations data by temporal attention, hierarchical attention and graph attention.

2. Iterative RNN-GNN architecture

In [24], Li et al. propose an LSTM Relational Graph Convolutional Network (LSTM-RGCN) model that predicts the overnight stock movement based on the correlation between stocks. This paper constructs a graph by converting each stock's news into a vector and calculating the relationship between each stock by using the cosine similarity. LSTM-RGCN is the first model that unitized connection among stocks to predict the movement of stocks that are not directly associated with news.

The LSTM-RGCN model, depicted in Figure 2(b), embodies an iterative RNN-GNN architecture. As illustrated in Figure 5, the process begins with LSTM encoding news data into vectors. Subsequently, the model merges the news vector with the node embedding to form the node vector. Finally, RGCN encodes the graph structure.:

$$N^{l+1} = \sigma\left(\sum_r D_r^{-\frac{1}{2}} A_r D_r^{-\frac{1}{2}} H^l W_r^l + W_h H^l\right) \tag{3}$$

where A_r is the adjacency matrix of relation r, $D_r^{-\frac{1}{2}} A_r D_r^{-\frac{1}{2}}$ is the normalized adjacency matrix. W_r^l is the learnable parameter matrix. W_h is the learnable parameter matrix for the node vector. The parameter matrices are shared across layers. H^l represents the hidden representations of all the nodes in the l-th layer. N^{l+1} is the aggregated neighbor information for the $(l+1)$-th layer.

Finally, the model predicts stock movement based on the node representation in the graph. $Sigmoid(\cdot)$ and $softmax$ functions are used as the activation functions. The cross-entropy is used as a loss function for this two-class classification task.

 ISBN: 978-1-68558-147-3

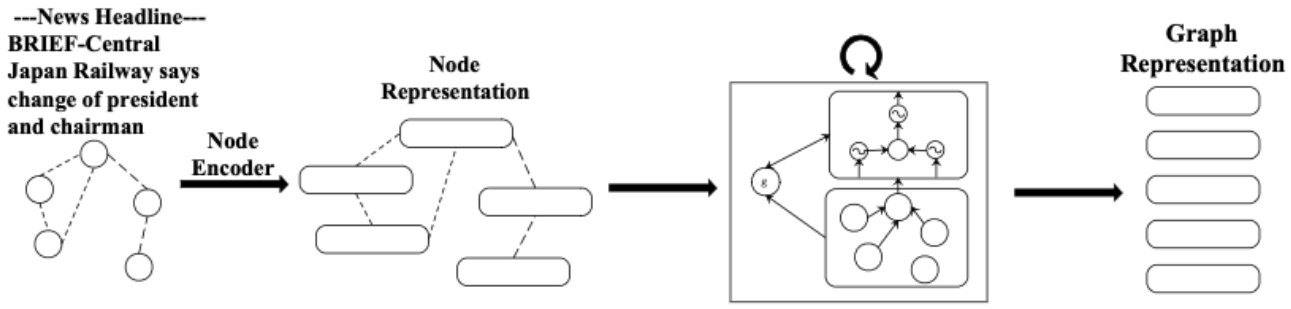

Fig. 5: LSTM-RGCN Model.

3. Parallel RNN ‖ GNN architecture

In their paper [23], Zhao et al. proposed a method called Dual Attention Networks to learn Stock Movement Prediction (DANSMP). This method leverages a market knowledge graph to model the relationships between stocks and make predictions about stock momentum. The graph comprises various types of information, including the relationships between companies and their executives.

DANSMP integrates three layers: stock sequential embedding, stock relational embedding, and prediction.

Initially, it merges technical and sentiment features using GRU for feature extraction. The relational layer uses dual attention networks for spillover signal representation, focusing on company-executive relationships via inter- and intra-class networks. Inter-class networks compare company and executive features, while intra-class networks assess same-type entity interactions.

Finally, embeddings are combined in a neural network for stock movement prediction. DANSMP's innovation lies in its parallel RNN-GNN structure and dual attention mechanism, enhancing market relationship analysis.

B. Dataset Feature

1. Data Information

The ICR-GCN dataset, sourced from Tushare API [29], comprises CSI 300 historical prices for listed companies from 29/04/2017 to 31/12/2017. It features five numeric attributes per company: open, close, high, low, and volume, utilizing seven days of historical data for input.

HATS dataset uses S&P 500 historical price data of listed companies between 08/02/2013 and 17/06/2019 (in total, 1174 trading days). For each company, there are three numeric features including open price, close price, and volume. The authors use historical price change rate $\boldsymbol{R}_i^t = \frac{\left(P_i^t - P_i^{t-1}\right)}{P_i^{t-1}}$ as model input, where P_i^t is the closing price at time t of a company i and P_i^{t-1} is the closing price at time $t-1$.

In the RSR-TGC model, data collection comprises three categories, starting with sequential price data from New York Stock Exchange (NYSE) and NASDAQ between February 1, 2013, and December 8, 2017. This dataset encompasses

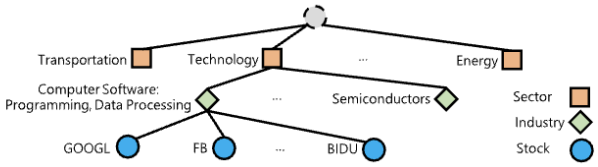

Fig. 6: Sector-industry relations and wiki company-based relations.

open, close, high, low, and volume metrics, using twenty-nine daily close prices for calculating (Moving Average) MA5, MA10, MA20, and MA30. These indicators, combined with the closing price, undergo normalization for model input.

The MAN-SF dataset used in this study was the StockNet dataset [30] which contains data on high-trade-volume stocks in the S&P 500 index in the NYSE and NASDAQ markets. The dataset is split into three parts: training, validation, and testing. The training data was used to train the MAN-SF model.

The DANSMP includes total of 185 stocks from the China Securities Index 300(CSI300E) and 73 stocks from the CSI100E are used to create two datasets which are collected from the China Securities Index (CSI). The market dataset includes historical price information (opening price, close price, highest price, lowest price and volume).

2. Stock Text Information

The second type in RSR-TGC model is sector-industry relations and the third type is Wiki relations such as supplier-consumer relations and ownership relations.

The LSTM-RGCN model also includes Financial news and market data from Tokyo Stock Exchange (TSE) from 1/1/2013 to 29/08/2018 from Reuters. There are two numeric features including open price and close price that are used to calculate the overnight movement. The formula for calculating the overnight movement is shown below:

$$\text{Movement} = \left(p_o^t - p_c^{t-1}\right) / p_c^{t-1} \tag{4}$$

where $p_o^t$is the open price of the current trading day and p_c^{t-1} is the close price of the previous trading day. Global Vectors for Word Representation (GloVe) [31] and Bidirectional Encoder

Representations from Transformers (BERT) [32] are used as word embedding on financial news to generate model input features.

In DANSMP model also includes the news dataset from four financial mainstream sites [33] [34] [35] [36].

C. Graph Construction Method

1. Correlation-based Graph Construction Methods

The ICR-GCN utilizes a data-driven graph construction method based on financial investment facts, using a graph of companies (stocks). This graph construction approach is designed to capture relevant relationships between companies. In this weighted graph, each node corresponds to a company, while the edges connecting the nodes represent the relationships between these companies. Furthermore, the weight assigned to each edge reflects the shareholding ratio between the connected companies.

2. Knowledge-based Graph Construction Methods

The HATS model constructs a heterogeneous graph from Wikidata to analyze relations between entities such as companies and persons. It simplifies this graph into a company-focused homogeneous one using a meta-path, representing companies as nodes connected by various relationship types, such as 'Owned by'.

The RSR-TGC model constructs its graph using sector-industry information, as illustrated in Figure 6, and company relations sourced from Wikidata. It categorizes stocks by industry and establishes connections between stocks based on first and second-order relations derived from Wikidata [37].

MAN-SF leverages first and second-order relations from Wikidata to map the S&P 500 index stocks' relationships, focusing on direct company connections.

DANSMP employs a Bi-typed Hybrid-relational Market Knowledge Graph (MKG) from Tushare API [29] data, featuring company and executive entities with both explicit (directly stated relationships) and implicit (inferred from attributes) relations. This approach enables a detailed analysis of company and executive interconnections for stock movement predictions, enriching market analysis and investment decisions.

3. Similarity-based Graph Construction Methods

In the LSTM-RGCN model, a stock correlation graph is built using historical prices, distinguishing relationships as positive or negative correlations. Connections between stock nodes are made if their similarity's absolute value meets a set threshold, enabling the model to understand stock interdependencies from historical prices.

IV. Discussion

In this section, we compare prior research, highlighting Table I's summary of model differences, including type, architecture, dataset, and data types. Table I shows the prevalent use of the RNN-GNN model, with three studies employing GCN-based GNNs (ICR-GCN, TGC, RGCN) and three using GAT-based models (HATS, MAN-SF, DANSMP). Dataset analysis reveals MAN-SF and HATS share a dataset; others vary, with sources ranging from Chinese to US and Japanese stock markets. Most datasets encompass price and news data. Graph-wise, DANSMP uniquely utilizes a heterogeneous graph; others use homogeneous ones, mostly constructed from Wikidata, except RGCN's price-based graphs.

We identify two key enhancement areas for stock prediction models:

1. Integrating transformers with GNN-based temporal models could improve accuracy by capturing long-range dependencies.

2. Developing dynamic spatial-temporal graphs with market data for transformer networks might refine models by leveraging complex spatial-temporal relationships.

These approaches suggest promising directions for refining stock market predictions, offering potential benefits for investors and analysts through advanced techniques and comprehensive market data utilization.

V. Conclusion

Stock market prediction is complex due to various impacting factors. This paper reviews six articles applying GNN to stock prediction, analyzing model architectures, data types, and graph construction methods.

Three model architecture patterns were identified, combining RNN and GNN, capturing temporal and relational stock data information. The primary data types were historical price, financial news, and financial indicators.

Graphs were constructed based on stock correlations, similarity measures, or network propagation techniques, capturing stock relationships and dependencies. Most models adopted the RNN-GNN framework, improving performance by leveraging the graph's structural information.

The review underscores the efficacy of GNN in stock prediction and the value of graph-based modelling with traditional sequential models. This paper provides an overview of GNN-based stock prediction, laying the foundation for future research.

References

[1] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the arima model," in *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*. IEEE, 2014, pp. 106–112.

[2] B. M. Awartani and V. Corradi, "Predicting the volatility of the s&p-500 stock index via garch models: the role of asymmetries," vol. 21, no. 1. Elsevier, 2005, pp. 167–183.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," vol. 9, no. 8. MIT Press, 1997, pp. 1735–1780.

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[5] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations*, Feb. 2017. [Online]. Available: http://arxiv.org/abs/1609.02907

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," vol. 20, no. 1. IEEE, 2008, pp. 61–80.

 ISBN: 978-1-68558-147-3

TABLE I: COMPARISON MODEL & METRICS BETWEEN ARTICLES

Framework	Model	Dataset	Graph relationships	AUX DATA
ICR-GCN	LSTM-GCN	CSI300	Stock-Stock	Financial investment fact from WIND
HATS	RNN-GAT	S&P500	Stock-Stock Stock-Owner	Wikidata
TGC	LSTM-TGC	NASDAQ NYSE	Stock-Stock	Wikidata
MAN-SF	GRU-GAT	S&P500	Stock-Stock Stock-Owner	Wikidata
RGCN	LSTM-RGCN	TPX500 TPX100	Stock-Stock	News
DANSMP	GRU-DAN	CSI100E CSI300E	Stock-Stock Stock-Owner Owner-Owner	News

[7] Y. Chen, Z. Wei, and X. Huang, "Incorporating corporation relationship via graph convolutional neural networks for stock price prediction," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1655–1658.

[8] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, 2020.

[9] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," *Expert Systems with Applications*, vol. 184, p. 115537, 2021.

[10] I. K. Nti, A. F. Adekoya, B. A. Weyori, and O. Nyarko-Boateng, "Applications of artificial intelligence in engineering and manufacturing: a systematic review," *Journal of Intelligent Manufacturing*, vol. 33, no. 6, pp. 1581–1601, 2022.

[11] A. W. Li and G. S. Bastos, "Stock market forecasting using deep learning and technical analysis: a systematic review," *IEEE access*, vol. 8, pp. 185 232–185 242, 2020.

[12] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Systems with Applications*, p. 116659, 2022.

[13] T. J. Strader, J. J. Rozycki, T. H. Root, and Y.-H. J. Huang, "Machine learning stock market prediction studies: review and research directions," *Journal of International Technology and Information Management*, vol. 28, no. 4, pp. 63–83, 2020.

[14] M. Nikou, G. Mansourfar, and J. Bagherzadeh, "Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms," *Intelligent Systems in Accounting, Finance and Management*, vol. 26, no. 4, pp. 164–174, 2019.

[15] M. Obthong, N. Tantisantiwong, W. Jeamwatthanachai, and G. Wills, "A survey on machine learning for stock price prediction: Algorithms and techniques," 2020. [Online]. Available: https://eprints.soton.ac.uk/437785/

[16] E. P. Lafortune and Y. D. Willems, "Bi-directional path tracing," 1993. [Online]. Available: https://masters.donntu.ru/2013/fknt/kalamitra/library/Bidirectional%20Path%20Tracing%201990.pdf

[17] P. Casanova, A. R. P. Lio, and Y. Bengio, "Graph attention networks," *ICLR. Petar Velickovic Guillem Cucurull Arantxa Casanova Adriana Romero Pietro Liò and Yoshua Bengio*, 2018.

[18] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[19] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," *Procedia computer science*, vol. 167, pp. 599–606, 2020.

[20] F. Li *et al.*, "Textual analysis of corporate disclosures: A survey of the literature," *Journal of accounting literature*, vol. 29, no. 1, pp. 143–165, 2010.

[21] J. Li and P. Meesad, "Combining sentiment analysis with socialization bias in social networks for stock market trend prediction," *International Journal of Computational Intelligence and Applications*, vol. 15, no. 01, p. 1650003, 2016.

[22] Y. Liu, Q. Zeng, J. Ordieres Meré, and H. Yang, "Anticipating stock market of the renowned companies: a knowledge graph approach," *Complexity*, vol. 2019, 2019.

[23] Y. Zhao *et al.*, "Stock movement prediction based on bi-typed and hybrid-relational market knowledge graph via dual attention networks," *arXiv preprint arXiv:2201.04965*, 2022.

[24] W. Li, R. Bao, K. Harimoto, D. Chen, J. Xu, and Q. Su, "Modeling the stock relation with graph network for overnight stock movement prediction." in *IJCAI*, vol. 20, 2020, pp. 4541–4547.

[25] R. Sawhney, S. Agarwal, A. Wadhwa, and R. Shah, "Deep attentive learning for stock movement prediction from social media text and company correlations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8415–8426.

[26] R. Kim, C. H. So, M. Jeong, S. Lee, J. Kim, and J. Kang, "Hats: A hierarchical graph attention network for stock movement prediction," *arXiv preprint arXiv:1908.07999*, 2019.

[27] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T.-S. Chua, "Temporal relational ranking for stock prediction," vol. 37, no. 2. ACM New York, NY, USA, 2019, pp. 1–30.

[28] H. Wang, S. Li, T. Wang, and J. Zheng, "Hierarchical adaptive temporal-relational modeling for stock trend prediction." in *IJCAI*, 2021, pp. 3691–3698.

[29] "Tushare." [Online]. Available: http://tushare.org/

[30] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1970–1979.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] "Wind." [Online]. Available: http://www.wind.com.cn/

[34] "Sina." [Online]. Available: http://www.sina.com/

[35] "Hexun." [Online]. Available: http://www.hexun.com/

[36] "Sohu." [Online]. Available: http://www.sohu.com/

[37] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," vol. 57, no. 10. ACM New York, NY, USA, 2014, pp. 78–85.

 ISBN: 978-1-68558-147-3

A Webtool for Hadith Authorship and Verification with Scholarly Expertise

Amina Daoud
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
e-mail: aminadaoud@hbku.edu.qa

Devi G. Kurup
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
e-mail: dkurup@hbku.edu.qa

Hamada R. H. Al-Absi*
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
e-mail: haalabsi@hbku.edu.qa

Jens Schneider
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
e-mail: jeschneider@hbku.edu.qa

Wajdi Zaghouani
College of Humanities and Social Sciences
Hamad Bin Khalifa University
Doha, Qatar
e-mail: wzaghouani@hbku.edu.qa

Saeed Mohd H. M. Al Marri
College of Shariah and Islamic Studies
Qatar University
Doha, Qatar
e-mail: saealmarri@qu.edu.qa

Younss Ait Mou*
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
e-mail: ymou@hbku.edu.qa

***Abstract*—Muslim scholars have collected Prophet Mohammed's teachings, sayings and actions several years after his death. These teachings have been collected into what is called "Hadith" in Islam, which is composed of several books. In recent years, digital databases have emerged based on the Hadith books; however, these databases lack validation by Muslim Hadith Scientists. This paper presents the development of a Webtool designed to validate Hadith collections, addressing the challenge of maintaining the integrity and accuracy of these essential Islamic texts in the digital era. By leveraging scholarly collaboration, the tool facilitates a thorough validation process for each Hadith and the chains of narrators. Feedback from Hadith Scientists has been instrumental in refining the tool's features, ensuring its academic accuracy and user-friendliness. The project signifies a crucial step towards preserving Islamic teachings and proposes a model for future technological advancements in the field of Islamic scholarship.**

***Index Terms*—Hadith; Hadith Scientist; Verification; Webtool.**

I. Introduction

A Hadith (which translates to speech or teaching) describes Prophet Mohammed's -Peace Be Upon Him (PBUH) - teachings, sayings, actions, or silent approvals [1]. The plural form of the word 'Hadith' is 'Ahadeeth,' or 'Hadiths' and these Ahadeeth have been observed/learned by the Prophet's companions throughout his life and were passed down through generations; however, it is essential to note that these Ahadeeth were only collected or documented and authenticated many years after the death of the Prophet PBUH [2]. These Ahadeeth are considered the second source of Islamic law; Muslims refer to them always besides the Holy Quran [3]. Each Hadith consists of two parts, the chain of narrators (Sanad/Isnad) which comes at the beginning and shows who have transmitted the Hadith, and the actual body of the Hadith (Matn/Matan), which is the actual teaching or saying or action of the Prophet PBUH. The proliferation of digitized Hadith collections online has brought concerns regarding their authenticity, accuracy, and reliability [2]. Many of these online databases lack scholarly validation, raising doubts among researchers and scholars [2]. For example, in this work, we used Hadith database of Sahih Bukhari from [4]. Hadith Scientists in our team have discovered and corrected multiple issues/errors in this database related to "Matan", "Sanad", the ordering of Hadiths, errors in chapters titles, etc. Table I shows some examples of these issues and how the Hadith Scientists corrected them.

Such issues or errors pose a significant challenge for those seeking to engage in rigorous academic or religious studies relying on these sources. To address these challenges and contribute to the field of Islamic scholarship, we aim to create a comprehensive and validated database of Hadith from the six major collections, including Sahih Bukhari, Sahih Muslim, Sunan Abu Dawood, Jami al-Tirmidhi, Sunan al-Sughra, and Sunan ibn Majah. With 114 chapters, 6,236 verses, and 157,935 words, the Quran provides teachings on daily matters, social dealings, and historical events. Advances in Natural Language Processing (NLP) enable computational techniques to facilitate Qur'anic research, offering new avenues for learning and understanding [5]. In our research, we employ NLP on Ahadeeth research, offering new techniques for learning and teaching. As an initial step for the research,

*Corresponding authors

 ISBN: 978-1-68558-147-3

TABLE I
EXAMPLES OF ISSUES/ERRORS FOUND IN THE ISLAMONLINE.NET HADITH DATABASE.

Type	Description
Matan	Hadith number 3535 was missing from the database.
Matan	Part of Hadith number 3757 was missing.
Sanad	In Hadith Number 3967, two of the narrators were substituted with well-known companions of the Prophet (Saad Bin Abi Waqas instead of Saad Bin Ibrahim and Abi Umama Al Bahili instead of Abu Umama Bin Sahl).
Chapters and Hadiths order	In the first few books, Hadiths were not ordered correctly (Note: this might be due to different editions of Sahih Bukhari) There were also errors in the ordering of chapters.
Books and Chapters	Some books have chapters with wrong title or without title. For example, in the Book of Marriage, there are four chapters, one without a title, one with the title: "Your prayers, your faith", another one with the title: "A Muslim is the one from whose hand Muslims are safe", and the fourth with the title "Food feeding". All titles were not related to marriage, yet the Hadiths enclosed in them are about marriage.

we create a comprehensive Hadith dataset through an expert authoring system. To ensure the accuracy and authenticity of each Hadith, we developed a Webtool which allows Hadith Scientists to manually validate the main text and the chain of narrators, correct any mistakes, and extract meaningful keywords for each Hadith. This tool allows several scholars to work in parallel. Ultimately, it provides a reliable and trustworthy resource for scholars, researchers, and the broader Muslim community. This contributes to the preservation and dissemination of authentic digital Hadiths for academic and religious purposes. At this stage, the Webtool has been developed and we started the validation of Sahih Al-Bukhari book. Noteworthy initiatives have paved the way for effective annotation processes. These include the development of interfaces like MADARi, a joint morphological annotation and spelling correction system for texts in Standard and Dialectal Arabic; and MANDIAC, a web-based annotation system designed for rapid manual diacritization of Standard Arabic text. These systems enable annotators to manage and organize the entire annotation process remotely and simultaneously [6]–[8]. To the best of our knowledge, there exists no previous work that has focused on validating Hadith digital collections, with Hadith Scientists going through each Hadith to ensure accuracy and integrity of the text.

The rest of this paper is organized as follows: Section II presents the interface design of the expert authoring and verification system; Section III presents a discussion on the web tool; and we present concluding remarks in Section IV.

II. EXPERT AUTHORING AND VERIFICATION SYSTEM INTERFACE DESIGN

A. User Requirements and Feedback

The primary users of this system are Hadith Scientists. These scientists are led by Professor Saeed Al-Marri from Qatar University, who is a principal investigator in this project and a renowned expert in Hadith science. Together with his team, they utilize their vast experience in Hadith science, and they use various books in the science of Hadith as well as specialized libraries, such as 'Al-Maktabah Al-Shamelah' – The Comprehensive Library. This library contains searchable books in different areas of the Islamic faith, including Hadith science. To facilitate the user interaction, an initial prototype was created to address their requirements which focused on:

- Verification of Hadith Authenticity – Hadith Scientists requested a feature to verify the displayed Hadith authenticity by entering comments on its "Matn" and its corresponding chain of narrators "Sanad".
- Correction of Hadith – The Hadith Scientists required a feature to correct any inaccuracies identified in the displayed Hadith.
- Displaying narrators' details – the users requested to include a functionality to access the details of each narrator in the chain, including information about their locations, number of Hadiths, number of students they had, etc.
- Additionally, the users also requested to save all the secondary names of narrators.

The prototype was used by the Hadith Scientists allowing them to interact with the system and to get their feedback which focused on the need to correct various discrepancies in Hadiths which included the following:

- Interface editability – Users requested for an editable interface, allowing them to correct the book names, chapter names, narrator names within the "Sanad", to enhance the accuracy of the verification process.
- Addition of collections, books, chapters, and Hadiths – Users expressed the interest of adding new collections, books, chapters and Hadith details instead of extracting data from online resources, which were not accurate.
- Refinement of Hadith Order – Users expressed the need of reordering the Hadiths through the interface for data accuracy.

These user requirements and feedback served as valuable insights, guiding the development and refinement of the verification system to meet the specific needs of Hadith Scientists.

 ISBN: 978-1-68558-147-3

B. System Architecture

The architecture of the interface for the expert authoring and verification system is designed to address the specific needs and feedback from Hadith Scientists. The system integrates various components to support data storage and retrieval of Hadith data in the form of user interaction and feedback incorporation. Data was initially sourced from [4], specifically Sahih Al-Bukhari data and is stored in MongoDB database. MongoDB is a popular NoSQL "Not only Structured Query Language" database, which uses a document-oriented data model that allows to store complex and unstructured data in a more natural way [9] [10]. During the initial phase of the data storage, we developed a robust structure to ensure an effective organization of data within the database. The database encompasses details about various elements such as Hadith collections, books, chapters, and specific details about each Hadith, including "Matn" and "Sanad". Other relevant information such as narrators' details, locations, etc. are stored effectively in their respective normalized collection within the database. As a first step of developing the interface, the first task was to create a login page to ensure the security of the system. This login page was designed to restrict access only to intended users, who are the Hadith Scientists. Creating a login page is an essential step in securing the application, guaranteeing that only authorized users have access. The login page is a simple and user-friendly interface that facilitates easy access for authorized users to interact and contribute to the verification system. Figure 1 shows the login page.

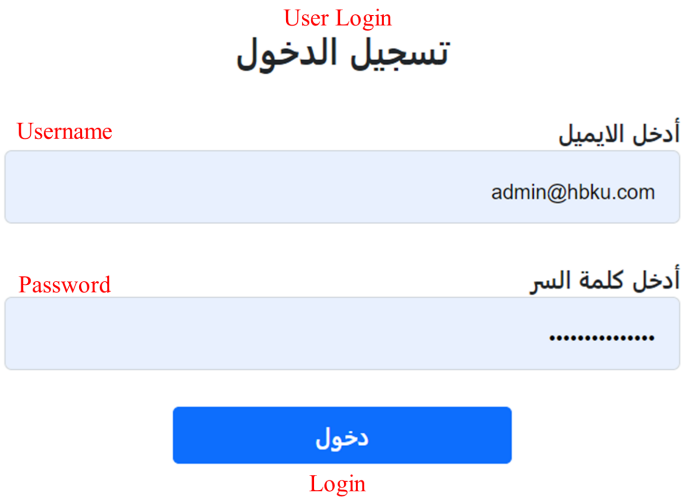

Fig. 1. Login page of the Webtool. English translation of all fields and functions are shown in red.

C. Core Features and Functionalities

The system's features and functionalities are comprehensively designed to cater to the needs of the primary users. A preliminary prototype was developed to enable the interaction between the primary users and the system. This prototype was developed based on the initial user requirements, focusing on the functionalities, such as Hadith authenticity verification, correction of inaccuracies in "Matns" and "Sanad", and the display of detailed narrator information. There is also a provision to add missing "Sanad" and "Moallakka" (*Moallakka Hadiths are those with incomplete initial chains of narrators, considered "suspended" in Islamic scholarship*) for the selected Hadith. By clicking on the displayed "Sanad" list, users can access corresponding detailed narrator information, including names, variations, locations, number of students, teachers, and associated Hadiths. Following user feedback, we incorporated additional features to the existing ones, encompassing the addition of new Hadith collections, books, chapters, and specific Hadith details. Moreover, the interface also includes the functionality to update Hadiths by employing a regular expression to separate "Sanad" and "Matn", enhancing the ease of extraction of those from Hadiths. Additionally, the system allows users to associate keywords with each Hadith, facilitating the information retrieval for research purposes. The system is designed in such a way that, for the initial use of the tool, it displays the first Hadith in the first chapter under the first book inside the first collection. As an update occurs, the subsequent openings of the tool will display the details of the last modified Hadith. Users can customize the display by selecting the collection followed by book, chapter and finally the desired Hadith for verification. The interface is thoughtfully organized into three blocks, aligned from right to left for Arabic letters. Each block can be maximized or minimized, enhancing user accessibility. The homepage provides functionalities for adding, updating, and deleting entries based on specific conditions. Additionally, the interface supports the reordering of Hadiths through a drag-and-drop method, allowing for the insertion of missing Hadiths while maintaining sequence numbers. To fortify security, the system incorporates measures like parameterized queries, to prevent injection attacks and ensuring the safety and validity of user inputs. The user interface is thoughtfully designed, featuring clear instructions for straightforward navigation. Figure 2 illustrates the main page of our Webtool application, which is segmented into three blocks:

- In the block on the right, there are three sections: collection, book, and chapter. Users have the option to select a collection, book, and chapter. Upon selection in the respective order, the Hadiths corresponding to the selections from these sections are displayed in their respective blocks. Additionally, each section provides an option to add new content, update existing content, and remove content if there are no subentries within their respective sections for collection, book, and chapter.
- The middle block displays all the Hadiths based on the selections made in the right block, with the first Hadith in the list highlighted, by default. Users have the option to choose which Hadiths to annotate, reorder Hadiths, change the chapters of Hadiths, add new Hadith, etc.
- The block on the left facilitates the users in editing the selected Hadith from the middle block. Users can rectify

 ISBN: 978-1-68558-147-3

Fig. 2. Webtool functionalities. English translation of all fields and functions are shown in red.

the mistakes in the selected Hadith (Matn), add and edit the chain of narrators (Sanad), Moallakka, including names and their order in the chain, add new narrator if they are not listed, and input comments related to the Sanad or the Matn. Furthermore, users can input relevant keywords for the selected Hadith.

The users can save the updates and navigate through Hadiths using the ”Next” and “Previous” buttons provided on the page without explicit user selection. Additionally, the users have the flexibility to expand and collapse each block according to their preference for enhanced accessibility. Figure 3 visually illustrates this functionality on the main page, with the third block expanded while the others collapsed, providing a clear navigation option for all users.

The system architecture supports an iterative development approach, allowing for continuous refinement based on user requirements and feedback. This comprehensive system architecture is tailored to meet the specific needs of Hadith Scientists, providing a robust and user-friendly environment for the verification and correction of Sahih Al-Bukhari data in this initial stage.

D. Tools and Technologies

The expert authoring and verification web interface system employs a comprehensive architecture, integrating Flask API [11]–[13] , MongoDB, HTML and jQuery to meet the specific needs of Hadith Scientists. The system’s structure is designed to facilitate data storage, retrieval, user interaction, and seamless incorporation of user feedback. The frontend of the web interface was developed with HTML and jQuery components, providing a user-friendly and interactive environment for Hadith Scientist to engage with the system. jQuery enables dynamic manipulation of the Document Object Model (DOM), enhancing the user experience during interactions with the interface. Figure 4 shows the architecture of the Webtool.

Flask is a lightweight and flexible web framework for Python and it is ideally suited for the development of web applications with simplicity and efficiency. Its modular design allows us to build a scalable and maintainable web application by integrating various extensions and libraries, as needed. Leveraging Flask built-in support for routing, request handling, and template rendering, the Webtool streamlines the development process while reducing boilerplate code. In our application, Flask API serves as an interface between the frontend and the database, handling HTTP requests from the frontend, processing data and orchestrating interactions with the MongoDB database. It communicates with the MongoDB database to retrieve, update and store data based on user interactions. MongoDB, a NoSQL database renowned for its flexibility, scalability, and performance in handling unstructured and semi-structured data, serves as the primary data storage solution. Its document-oriented model enables the storage of data in JSON-like documents, making it suitable for storing diverse and evolving data structures. MongoDB’s schemaless design eliminates the need for a predefined schema, facilitating easy modification of data models. MongoDB is used for data storage in our application, storing data in a document-oriented format. It is queried and updated by the Flask API to provide the required information to the frontend. By combining Flask, MongoDB, and HTML, we have developed a robust and feature-rich Webtool that leverage the strengths of each technology. Flask provides the backend infrastructure and routing capabilities, MongoDB offers a flexible and scalable data storage solution, and HTML

 ISBN: 978-1-68558-147-3

Fig. 3. Hadith block after expansion. English translation of all fields and functions are shown in red.

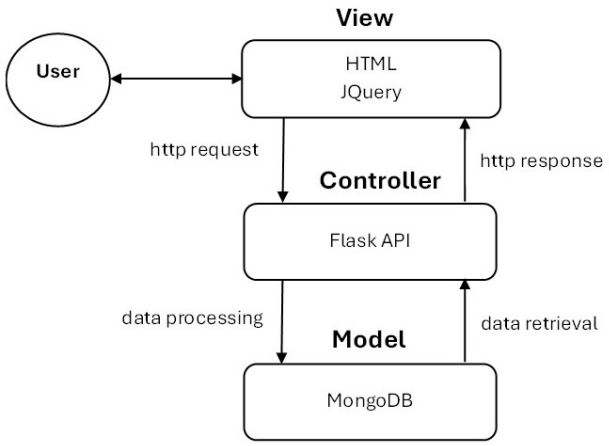

Fig. 4. Webtool Architecture.

enables the creation of intuitive and user-friendly interface. To further enhance the organization and maintainability of our codebase, we have adopted Flask Blueprints [13]. These blueprints serve as a powerful tool, allowing us to organize our application into separate components, each handling specific functionalities. This modular approach facilitates code reuse and scalability, making it particularly beneficial for large-scale applications where maintaining a well-organized codebase is crucial. Specifically, we utilized blueprints for functionalities such as user login, the main page, where the core functionalities are maintained, and the display of narrator details. This ensures that our code remains modular, scalable, and easily maintainable, contributing to the overall efficiency and robustness of our application. By combining these tools and technologies, the expert authoring and verification system is built to meet the specific needs of our primary users, providing a robust, secure, and user-friendly platform for the verification and correction of Hadiths from six collections.

III. Discussion

The development of a Webtool aimed at authenticating Hadith collections is a significant leap forward in the intersection of Islamic scholarship and digital technology. This work acknowledges the critical challenge of preserving the authenticity and accuracy of Hadith, an essential component of Islamic faith and jurisprudence, in the vast and often unregulated digital landscape. By enabling scholarly collaboration, the tool facilitates a thorough validation process, ensuring each Hadith's text and its chain of narrators are carefully examined. This collaborative approach not only streamlines the authentication process but also provides access to validated Hadiths, empowering scholars and the broader Muslim community with reliable resources. Moreover, the integration of feedback from Hadith scholars into the tool's design underscores a dynamic and responsive development process. This feedback is critical for refining the tool's functionalities, such as verification of authenticity, correction of inaccuracies, and the addition of comprehensive narrator details. Such features address the needs of Hadith scholars, ensuring the tool remains both user-friendly and academically accurate. This responsiveness to scholarly input highlights the project's commitment to serving the needs of Islamic academia and contributes to its potential as a cornerstone resource for Hadith studies.

 ISBN: 978-1-68558-147-3

IV. Conclusion

In this work, we presented a Webtool to authenticate and verify Hadiths. This work extends beyond the immediate benefits of improved access to authenticated Hadiths. It sets a precedent for future works in the digital preservation and study of Islamic texts, suggesting pathways for integrating technological solutions with traditional Islamic scholarship. The project's success could inspire further innovations in the field, including the development of similar tools for other Islamic texts and the adoption of advanced technologies for scholarly research. As such, this Webtool represents not just a solution to a current challenge but a steppingstone towards a future where technology and tradition blend to advance Islamic scholarship.

Acknowledgment

This work has been supported by Qatar National Research Fund (QNRF), now known as the Qatar Research Development and Innovation (QRDI) Council, under grant number (NPRP13S-0129-200194).

References

[1] A. Ramzy *et al.*, "Hadiths classification using a novel author-based hadith classification dataset (abcd)," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 141, 2023.

[2] S. Hakak *et al.*, "Digital hadith authentication: Recent advances, open challenges, and future directions," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 6, p. e3977, 2022.

[3] F. Binbeshr, A. Kamsin, and M. Mohammed, "A systematic review on hadith authentication and classification methods," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 2, pp. 1–17, 2021.

[4] IslamOnline, https://islamonline.net/, accessed: 15/12/2023.

[5] M. H. Bashir *et al.*, "Arabic natural language processing for qur'anic research: a systematic review," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6801–6854, 2023.

[6] O. Obeid *et al.*, "Madari: A web interface for joint arabic morphological annotation and spelling correction," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 2616–2622.

[7] ——, "Mandiac: A web-based annotation system for manual arabic diacritization," in *The 2nd Workshop on Arabic Corpora and Processing Tools*, 2016, pp. 16–21.

[8] ——, "A web-based annotation framework for large-scale text correction," in *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, 2013, pp. 1–4.

[9] B. Jose and S. Abraham, "Exploring the merits of nosql: A study based on mongodb," in *2017 International Conference on Networks and Advances in Computational Technologies (NetACT)*. IEEE, 2017, pp. 266–271.

[10] A. Utka *et al.*, "Berri corpus manager: A corpus analysis tool using mongodb technology," in *Human Language Technologies–The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020*, vol. 328. IOS Press, 2020, pp. 166–173.

[11] P. Smyth, *Creating web APIs with Python and Flask. Programming Historian 7*. ProgHist Ltd, 2018.

[12] D. Gaspar and J. Stouffer, *Mastering Flask Web Development: Build Enterprise-grade, Scalable Python Web Applications*. Packt Publishing Ltd, 2018.

[13] J. Perras, *Flask Blueprints*. Packt Publishing Ltd, 2015.

 ISBN: 978-1-68558-147-3