ISSN: 3049-1479 (Online)

A Comprehensive Study of Big Data Analytics and Social Media

Hiteshkumar Omprakash Maidh[1], Mahammad Idrish I. Sandhi[2]

Research Scholar, Computer Science, Sankalchand Patel University, Visnagar, India[1]
Associate Professor & Head, Department of Computer Application (MCA), Sankalchand Patel University, Visnagar, India [2]

hitesh_maidh@yahoo.co.in[1], idrish.mca@gmail.com[2]

Abstract: The exponential growth of social media platforms has led to an unprecedented surge in data generation, offering vast opportunities for research, analysis, and insight generation. This literature review synthesizes existing studies on the implementation of big data analytics on social media data. It delves into various aspects including methodologies, tools, challenges, and emerging trends.

The review begins by exploring the methodologies employed in analyzing social media data, ranging from traditional statistical methods to advanced computational methods of machine learning. It highlights the importance of selecting appropriate methodologies based on the research objectives and characteristics of the data.

Furthermore, the review addresses the diverse range of tools and platforms available for big data analysis in social media data, including open-source frameworks, commercial software, and custom-built solutions. It examines the functionalities, scalability, and usability of these tools, offering insights into their suitability for different research contexts.

Keywords: Social Network Analytics, Big Data Analytics, Big Data, Traditional Databases, Systematic Literature Review.

I. INTRODUCTION

The advent of big data has ushered in a transformative era across academic, industrial, and public sectors, enabling unprecedented insights through the analysis of massive, complex datasets. Big data analytics, characterized by the "3Vs"—volume, velocity, and variety—has become an indispensable tool for decision-making, predictive modeling, and behavioral analysis (Bach et al., 2019; Luckow et al., 2016). The proliferation of social media, sensor networks, financial transactions, and video surveillance systems has contributed to a deluge of real-time data streams that, when analyzed effectively, can significantly improve operations, planning, and strategic responses across domains (Felt, 2016; Tsou, 2015).

In the financial sector, text mining techniques have emerged as a dominant method for processing unstructured textual data to detect patterns, assess risks, and inform investment decisions (Bach et al., 2019). Meanwhile, deep learning technologies have found wide application in intelligent video surveillance, particularly in analyzing crowd behavior and detecting anomalies in real-time (Sreenu & Durai, 2019). Social media platforms, with their massive user-generated content, have become a fertile ground for sentiment analysis, particularly in monitoring public discourse during crises like the COVID-19 pandemic (Zhu et al., 2020). These platforms also play a pivotal role in urban management and emergency event detection, where crowdsourced data can enhance situational awareness (Xu et al., 2020).

Despite these promising applications, the deployment of big data analytics raises several concerns. For example, the growing influence of fake news and misinformation challenges the reliability of data-driven insights in media landscapes (Vargo et al., 2018). Moreover, biases in data collection and representation can lead to the exclusion of marginalized voices, reducing the fairness and inclusivity of data-driven outcomes (Hargittai, 2018). The technical challenge of

https://doi.org/10.63766/spujstmr.24.000027

addressing non-functional requirements—such as scalability, security, and real-time performance—adds further complexity to the development of robust big data systems (Rahman & Reza, 2020). Furthermore, a systematic review by Abkenar et al. (2020) emphasizes the fragmented nature of big data research in social media, noting a lack of integration across techniques and application domains.

Given the interdisciplinary nature and rapid evolution of big data analytics, a comprehensive review is necessary to consolidate existing knowledge, identify current trends, and highlight gaps in the literature. This review aims to synthesize findings from recent studies across key domains—finance, surveillance, public health, and social media—to present an integrated perspective on the state of big data analytics. It also explores the methodological innovations, practical applications, and ethical implications that shape the current landscape, providing a foundation for future research and development.

II. LITERATURE REVIEW

The expansion of big data technologies has driven a surge in research and development across multiple domains. This literature review synthesizes existing studies to explore how big data analytics has been applied in finance, social media, surveillance, urban emergency management, and other sectors, while also examining technical challenges and ethical concerns associated with big data systems.

A. Big Data Applications in the Financial Sector

The financial industry has been one of the earliest adopters of big data analytics, primarily for risk assessment, market prediction, fraud detection, and customer segmentation. Bach et al. (2019) provided a comprehensive literature review on the role of text mining within the financial sector, emphasizing its use in analyzing unstructured data such as financial news, earnings reports, and social media. Their review highlights how machine learning algorithms enhance forecasting and decision-making capabilities by extracting actionable insights from massive text corpora. However, the authors also note critical challenges, including data quality, semantic ambiguity, and the need for domain-specific ontologies to improve classification and clustering tasks.

B. Social Media Analytics and Public Sentiment

Social media has emerged as a rich source of big data, offering insights into public sentiment, behavior, and societal trends. Felt (2016) examined how researchers in the social sciences leverage big data from platforms like Twitter and Facebook to explore issues such as political discourse, health communication, and crisis response. Despite its promise, Felt cautions that social media data may reflect the biases of digitally connected populations, thus potentially omitting marginalized voices.

Zhu et al. (2020) explored spatiotemporal trends in public sentiment by analyzing COVID-19-related discussions on Chinese social media platforms. Their findings demonstrated how big data techniques—particularly sentiment analysis and spatial mapping—can capture evolving public attitudes and support policy-making during health crises. However, the authors also acknowledged difficulties in dealing with noise and misinformation inherent in social media data.

Similarly, Tsou (2015) highlighted the geospatial potential of social media big data, suggesting its utility in tracking disease outbreaks, natural disasters, and migration patterns. Mapping this data, however, requires integration with geographic information systems (GIS) and robust filtering techniques to ensure reliability.

C. Big Data and Fake News Detection

Big data analytics also plays a dual role in media environments: while it can aid in understanding information dissemination, it can also inadvertently contribute to the spread of misinformation. Vargo et al. (2018) analyzed online news content from 2014 to 2016 and found that fake news gained substantial visibility, often matching or exceeding that of mainstream sources. Their work illustrates how algorithms designed to maximize engagement can be exploited to amplify misleading content. This has serious implications for democracy, public trust, and social cohesion, making misinformation detection a growing research priority.

D. Urban Analytics and Emergency Event Response

In the context of smart cities and disaster management, big data is increasingly used to improve emergency response and urban planning. Xu et al. (2020) proposed a crowd sourcing-based framework that utilizes real-time social media posts to describe and respond to urban emergencies. Their system integrates cloud computing with geospatial data to identify critical incidents more effectively. While this approach shows promise in enhancing situational awareness, challenges include data validation, real-time processing, and managing large-scale heterogeneous inputs.

E. Surveillance and Public Safety Through Deep Learning

Public safety applications, particularly in intelligent video surveillance, have seen rapid advancements through the integration of deep learning with big data. Sreenu and Durai (2019) provided an extensive review of deep learning models used for crowd analysis and behavior recognition. Convolution Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Recurrent Neural Networks (RNNs) are highlighted as effective tools for identifying suspicious behavior and managing large public gatherings. While accuracy and automation are improving, ethical questions concerning privacy and surveillance culture remain largely unresolved.

F. Industrial Applications: Automotive Sector

Big data and deep learning are also revolutionizing the automotive industry, particularly in areas such as autonomous driving, predictive maintenance, and driver behavior analysis. Luckow et al. (2016) discussed several applications and tools used in the automotive domain, emphasizing the use of distributed computing frameworks such as Hadoop and Spark. Their work underlines the importance of high-performance computing infrastructure to process the massive datasets generated by connected vehicles and in-vehicle sensors.

G. Technical Challenges in Big Data Systems

While the functional capabilities of big data systems have grown significantly, non-functional requirements such as performance, scalability, security, and fault tolerance remain under-researched. Rahman and Reza (2020) conducted a systematic mapping study to examine how these requirements are addressed in current big data projects. Their results reveal gaps in ensuring end-to-end system robustness and point to the need for standardized development practices.

Additionally, Abkenar et al. (2020) conducted a systematic review of big data analytics in social media, highlighting fragmented methodologies and limited cross-domain integration. They call for a unified framework that bridges data acquisition, storage, processing, and visualization—especially when handling social media content that varies widely in quality, language, and format.

H. Ethical and Methodological Concerns

The power of big data is tempered by growing ethical and methodological concerns. Hargittai (2018) raised alarms about bias in big data, noting that individuals who are not active online or whose data is excluded due to platform limitations are often left out of analyses. This can lead to skewed policy decisions and academic interpretations. Ensuring fairness, transparency, and inclusivity in big data research is therefore critical for maintaining social trust and achieving equitable outcomes.

III. RESEARCH METHODOLOGY

This review adopts a systematic literature review (SLR) methodology to explore the applications, trends, challenges, and future directions of big data analytics across multiple domains. The methodology follows established guidelines for review studies, ensuring that the selection and analysis of literature are rigorous, transparent, and replicable (Abkenar et al., 2020; Rahman & Reza, 2020).

A. Research Objectives

The main objectives of this review are to:

- Identify key application areas of big data analytics in sectors such as finance, social media, urban planning, surveillance, and public health.
- Examine the technological frameworks, analytical methods, and deep learning techniques used across studies.
- Explore the challenges associated with data quality, misinformation, bias, scalability, and non-functional requirements.
- Synthesize ethical and methodological concerns related to the use of big data systems.

B. Literature Search Strategy

A comprehensive search was conducted across scholarly databases including IEEE Explore, Science Direct, Springer Link, Scopus, Web of Science, and Google Scholar. The search period was limited to the years 2015 to 2020, aligning with the timeline of major technological advances in big data and artificial intelligence. The following keywords and Boolean operators were used:

("big data analytics" OR "big data applications") AND ("finance" OR "social media" OR "surveillance" OR "urban planning" OR "COVID-19" OR "crowd analysis" OR "deep learning" OR "text mining")

This search yielded a total of 92 initial publications.

C. Inclusion and Exclusion Criteria

To ensure relevance and quality, studies were evaluated using predefined inclusion and exclusion criteria:

Inclusion Criteria:

• Peer-reviewed journal articles and conference papers.

• Studies focusing on the use of big data analytics in real-world applications.

• Papers that incorporate advanced techniques such as machine learning, deep learning, or natural language processing

(NLP).

• Publications written in English between 2015 and 2020.

Exclusion Criteria:

• Non-peer-reviewed sources (e.g., blogs, white papers).

• Articles focused solely on hardware infrastructure or unrelated IT domains.

• Duplicate publications or those without accessible full text.

After applying these filters, 11 key publications were retained for in-depth analysis. These include comprehensive reviews (Abkenar et al., 2020; Bach et al., 2019), domain-specific studies (Zhu et al., 2020; Sreenu & Durai, 2019), and papers highlighting cross-domain challenges (Hargittai, 2018; Rahman & Reza, 2020).

D. Thematic Analysis and Categorization

Each selected study was reviewed and categorized according to its primary research domain and contribution type. The following thematic categories were used:

• Financial analytics (e.g., text mining, fraud detection) – Bach et al. (2019)

• Social media sentiment and misinformation – Felt (2016); Zhu et al. (2020); Vargo et al. (2018)

• Surveillance and public safety – Sreenu & Durai (2019)

- Urban emergency and crowd response – Xu et al. (2020)
- Industry-specific applications (e.g., automotive) – Luckow et al. (2016)
- Technical architecture and non-functional requirements – Rahman & Reza (2020)
- Bias and ethical concerns – Hargittai (2018)

Each paper was coded for its methods (qualitative, quantitative, computational), data sources (social media, sensor data, news articles), analytical techniques (machine learning, NLP, GIS), and reported limitations.

E. Analytical Framework

Following the thematic classification, a comparative matrix was developed to map the methodological approaches, tools used (e.g., Hadoop, Spark, Tensor Flow), and outcomes. This matrix helped in identifying patterns, similarities, and divergences across domains.

Moreover, this review employs a narrative synthesis approach to interpret results and draw interconnections between technical implementations and their social implications, similar to the approach used by Abkenar et al. (2020) and Tsou (2015).

IV. KNOWLEDGE GAPS AND FUTURE RESEARCH DIRECTIONS

While big data analytics has gained considerable traction across various sectors, including finance, public health, surveillance, social media, and smart cities, a review of the literature reveals several persistent knowledge gaps. These gaps limit the effectiveness, scalability, and ethical use of big data and signal opportunities for future research to advance the field both methodologically and practically.

A. Lack of Standardized Methodological Frameworks

One of the most consistent gaps across the literature is the lack of standardization in analytical methodologies. Studies often employ different data sources, preprocessing techniques, and evaluation metrics, making cross-comparison and reproducibility difficult (Abkenar et al., 2020). For instance, social media sentiment analysis studies use diverse sentiment lexicons and machine learning algorithms with varying levels of transparency, leading to inconsistent findings (Zhu et al., 2020; Felt, 2016).

Future Direction:

There is a need to establish standardized protocols and benchmarks for data collection, feature engineering, model evaluation, and validation in big data research. Developing open-source libraries and shared datasets across domains can facilitate comparative analysis and methodological rigor.

B. Inadequate Handling of Data Quality and Real-Time Constraints

A major technical challenge identified in current literature is managing data quality, especially from unstructured or user-generated sources like social media and video surveillance (Xu et al., 2020; Sreenu & Durai, 2019). Noise, misinformation, and incomplete data can reduce the reliability of analytical outcomes. Moreover, real-time applications such as emergency response or crowd monitoring demand ultra-low latency processing, which is often not addressed in conventional big data frameworks.

Future Direction:

Future research should invest in advanced data-cleaning algorithms, context-aware filtering, and real-time processing architectures using tools like Apache Flink or edge computing. Techniques such as hybrid human–AI systems may also enhance accuracy in mission-critical scenarios by combining machine efficiency with human judgment.

C. Underexplored Ethical Implications and Algorithmic Bias

Many studies acknowledge the importance of ethics but fall short of implementing systematic frameworks to mitigate algorithmic bias and uphold data justice (Hargittai, 2018; Vargo et al., 2018). The exclusion of certain demographic groups from social media datasets can lead to skewed models, while opaque algorithms may reinforce discriminatory outcomes in areas like credit scoring or policing.

Future Direction:

Researchers must embed Fairness, Accountability, and Transparency (FAT) principles into the development cycle of big data systems. This includes bias detection tools, interpretable AI models, and frameworks for ethical auditing. In addition, inclusive data collection practices should be prioritized to represent marginalized and underrepresented communities.

D. Limited Focus on Non-Functional Requirements in System Design

Big data research often emphasizes functional performance (e.g., accuracy, precision) while neglecting non-functional system requirements such as scalability, fault tolerance, and system security (Rahman & Reza, 2020). In large-scale deployments, especially in sectors like healthcare or urban infrastructure, these factors are critical for long-term viability.

Future Direction:

Future work should emphasize architectural innovations that address non-functional requirements. For instance, combining cloud computing with block chain can enhance data security, transparency, and system resilience. The use of edge computing and distributed systems can also improve scalability and reduce latency.

E. Weak Integration Across Disciplines and Domains

Despite the interdisciplinary potential of big data, current research tends to be siloed, with little integration between technical and domain-specific expertise. For example, studies on deep learning in surveillance (Sreenu & Durai, 2019) or finance (Bach et al., 2019) often lack input from ethics, behavioral science, or public policy scholars, which could provide valuable perspectives.

Future Direction:

Promoting interdisciplinary collaborations is essential for the next generation of big data solutions. Joint research initiatives involving computer scientists, sociologists, policy makers, and ethicists can produce more holistic and responsible systems. Funding bodies should support collaborative research grants that mandate cross-domain partnerships.

F. Inattention to Longitudinal and Context-Aware Analytics

Most current studies adopt a snapshot-based approach, focusing on specific events or datasets without accounting for temporal trends or contextual variables (Zhu et al., 2020). This limitation hinders the ability to detect patterns of long-term change or regional variation.

Future Direction:

Future research should prioritize longitudinal studies that track data over extended periods to uncover sustained behavioral, financial, or social dynamics. Moreover, integrating context-aware modeling—accounting for geography, culture, time, and policy environments—can significantly improve the accuracy and relevance of predictive analytics.

G. Limited Exploration of Explainable and Interpretable Models

While many studies use complex models like deep learning, few focus on making these models interpretable to end-users or decision-makers. This reduces transparency and hinders trust in automated systems, especially in high-stakes environments like finance, law enforcement, or healthcare.

Future Direction:

Future work should explore Explainable AI (XAI) methods to improve model interpretability and user confidence. Techniques such as SHAP (SHapley Additive explanations), LIME (Local Interpretable Model-agnostic Explanations), and rule-based modeling can help visualize and explain the decision-making processes of black-box models.

H. Lack of Evaluation in Real-World Settings

Many studies are limited to simulations or small-scale experiments and are rarely tested in real-world environments. This gap reduces the generalizability and scalability of proposed solutions.

Future Direction:

Future research should include real-world pilot implementations, preferably in collaboration with industry or government agencies. This would enable the validation of models under operational constraints and offer practical feedback to refine systems.

TABLE I

SUMMARY OF RESEARCH OPPORTUNITIES

Knowledge Gap	Future Research Direction
Methodological inconsistency	Develop standard frameworks and benchmarks
Data noise and real-time processing	Advance filtering and edge computing architectures
Ethical concerns and bias	Implement fairness-aware algorithms and inclusive data policies
Non-functional system requirements	Design secure, scalable architectures (e.g., block chain, cloud-edge)
Disciplinary silos	Promote interdisciplinary research and co-authored models
Snapshot-based analysis	Conduct longitudinal and context-aware studies
Black-box modeling	Develop and integrate Explainable AI tools
Lack of real-world testing	Run pilot studies in real environments with stakeholders

V. CONCLUSION

Big data analytics has evolved into a transformative paradigm with far-reaching implications across sectors such as finance, healthcare, public safety, urban planning, and social media. This review paper synthesized findings from diverse studies to illustrate how big data technologies—combined with tools like machine learning, natural language processing, deep learning, and cloud computing—are being leveraged to extract actionable insights from massive, complex datasets.

The reviewed literature demonstrates significant progress in applying big data for practical outcomes, such as fraud detection in finance (Bach et al., 2019), intelligent video surveillance (Sreenu & Durai, 2019), real-time emergency management (Xu et al., 2020), and sentiment tracking during pandemics (Zhu et al., 2020). These applications reflect not only the growing computational capabilities but also the creative adaptation of algorithms to tackle real-world challenges.

However, this progress is counterbalanced by a series of persistent challenges and knowledge gaps that limit the scalability, fairness, and generalizability of big data solutions. Methodological inconsistencies, data quality issues, ethical concerns, underdeveloped system architecture, and disciplinary silos are critical barriers that must be addressed. For instance, while many studies show the power of deep learning, few explore the explainability of models—an essential factor for trust and adoption in high-stakes domains (Hargittai, 2018; Abkenar et al., 2020). Likewise, the neglect of non-functional system requirements such as scalability and resilience (Rahman & Reza, 2020) threatens the robustness of real-world implementations.

Furthermore, the siloed nature of research has resulted in fragmented knowledge. Despite the interdisciplinary potential of big data, integration across domains—particularly between technical and social sciences—remains limited. Bridging this gap is vital for ensuring that big data technologies are not only efficient but also socially and ethically responsible.

Looking ahead, the field stands at a pivotal moment. The next phase of big data research must focus on developing standardized, explainable, scalable, and ethically sound systems. This includes fostering interdisciplinary collaborations, adopting FAIR (Findable, Accessible, Interoperable, Reusable) data principles, and embedding ethical oversight throughout the development lifecycle.

In summary, while big data analytics holds immense promise, realizing its full potential requires addressing critical gaps in methodology, ethics, inclusivity, and systems design. Through deliberate, collaborative, and reflective research practices, future work can contribute to building a more intelligent, equitable, and data-driven society.

REFERENCES

[1] Bach, M. P.., Krstic, Z.., Seljan, S.., & Turulja, L.. (2019). *Text Mining for Big Data Analysis in Financial Sector: A Literature Review. Sustainability* . http://doi.org/10.3390/SU11051277

[2] Sreenu, G.., & Durai, M. A. Saleem. (2019). Intelligent *video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data* , 6 . http://doi.org/10.1186/s40537-019-0212-5

[3] Zhu, Bangren., Zheng, Xinqi., Liu, Haiyan., Li, Jiayang., & Wang, Peipei. (2020*). Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics*. Chaos, Solitons, and Fractals , 140 , 110123 - 110123 . http://doi.org/10.1016/j.chaos.2020.110123

[4] Vargo, Chris J.., Guo, Lei., & Amazeen, Michelle A.. (2018). *The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016*. New Media & Society , 20 , 2028 - 2049 . http://doi.org/10.1177/1461444817712086

[5] Felt, Mylynn. (2016). *Social media and the social sciences: How researchers employ Big Data analytics. Big Data & Society* , 3 . http://doi.org/10.1177/2053951716645828

[6] Xu, Zheng., Liu, Yunhuai., Yen, N.., Mei, Lin., Luo, Xiangfeng., Wei, Xiao., & Hu, Chuanping. (2020). *Crowdsourcing Based Description of Urban Emergency Events Using Social Media Big Data*. IEEE Transactions on Cloud Computing , 8 , 387-397 . http://doi.org/10.1109/TCC.2016.2517638

[7] Luckow, André., Cook, M.., Ashcraft, Nathan., Weill, Edwin., Djerekarov, Emil., & Vorster, Bennie. (2016). *Deep learning in the automotive industry: Applications and tools*. 2016 IEEE International Conference on Big Data (Big Data) , 3759-3768 . http://doi.org/10.1109/BigData.2016.7841045

[8] Tsou, Ming-Hsiang. (2015). *Research challenges and opportunities in mapping social media and Big Data.* Cartography and Geographic Information Science , 42 , 70 - 74 . http://doi.org/10.1080/15230406.2015.1059251.

[9] Hargittai, E.. (2018). *Potential Biases in Big Data: Omitted Voices on Social Media*. Social Science Computer Review , 38 , 10 - 24 . http://doi.org/10.1177/0894439318788322

[10]Md. Saifur Rahman, Hassan Reza (2020). *Systematic Mapping Study of Non-Functional Requirements in Big Data System*. http://doi.org/10.1109/EIT48999.2020.9208288

[11]Sepideh Bazzaz Abkenar, Mostafa Haghi Kashani, Ebrahim Mahdipour, Seyed Mahdi Jameii (2020). *Big data analytics meets social media: A systematic review of techniques, open issues, and future directions*. https://doi.org/10.1016/j.tele.2020.101517