Optimizing social media analytics with the data quality enhancement and analytics framework for superior data quality

B. Karthick[1,2], T. Meyyappan[1]
[1]Department of Computer Science, Alagappa University, Karaikudi, India
[2]Department of Computer Science, Syed Hameedha Arts and Science College, Kilakarai, India

Article Info

Article history:

Received Aug 18, 2024
Revised Apr 8, 2025
Accepted Jun 10, 2025

Keywords:

Big data
Data quality
Random forest
Social media analysis
Support vector machine
Tumblr

ABSTRACT

This paper introduces the data quality enhancement and analytics (DQEA) framework to enhance data quality in social media analytics through machine learning (ML) algorithms. The efficacy of the framework is validated through features tested against human coders on Amazon Mechanical Turk, achieving an inter-coder reliability score of 0.85, indicating high agreement. Furthermore, two case studies with a large social media dataset from Tumblr were conducted to demonstrate the effectiveness of the proposed content features. In the first case study, the DQEA framework reduced data noise by 30% and bias by 25%, while increasing completeness by 20%. In the second case study, the framework improved data consistency by 35% and overall data quality score by 28%. Comparative analysis with state-of-the-art models, including random forest and support vector machines (SVM), showed significant improvements in data reliability and decision-making accuracy. Specifically, the DQEA framework outperformed the random forest model by 15% in accuracy and 20% in true positive rate, and the SVM model by 10% in error rate reduction and 18% in reliability. The results underscore the potential of advanced data analytics tools in transforming social media data into a valuable asset for organizations, highlighting the practical implications and future research directions in this domain.

Corresponding Author:

B. Karhtick
Department of Computer Science, Alagappa University
Karaikudi-630003, Tamil Nadu, India
Email: bkarthick1980@gmail.com

1. INTRODUCTION

The proliferation of social media platforms in recent years has transformed the way individuals and organizations communicate, share information, and engage with their audiences. Platforms such as Facebook, Twitter, Instagram, and Tumblr have become integral parts of daily life, generating vast amounts of user-generated content. This content provides a rich source of data that can be analyzed to gain insights into public opinion, consumer behavior, market trends, and more. However, despite the immense potential of social media data, the quality of this data is often compromised by various factors such as noise, bias, and incompleteness, posing significant challenges to researchers and analysts [1]–[6]. Noise in social media data refers to irrelevant or extraneous information that does not contribute to meaningful analysis. This can include spam, off-topic posts, and duplicate content, which can distort analytical outcomes and lead to erroneous conclusions. Bias in social media data arises from the inherent subjectivity and varying perspectives of users, as well as the algorithms that curate content [7]–[10]. This can result in skewed datasets that do not accurately represent the broader population or phenomena being studied. Incompleteness, another critical issue, occurs when datasets lack sufficient data points or have missing information, leading to gaps in analysis and unreliable results. Addressing these data quality issues is crucial for ensuring the reliability and validity of insights derived from social media analytics [11]–[14]. Traditional approaches to enhancing data quality, such as business decision management systems (BDMS), have been employed to mitigate these challenges. However, these methods often fall short due to their reliance on predefined rules and manual interventions, which may not scale effectively with the dynamic and voluminous nature of social media data [15]–[18]. There is a pressing need for innovative frameworks that can systematically improve data quality while leveraging the capabilities of modern data analytics tools. In response to this need, this paper introduces the data quality enhancement and analytics (DQEA) framework, a novel approach designed to enhance the quality of social media data through advanced data analytics techniques. Unlike traditional methods, the DQEA framework utilizes a combination of automated data processing, integration, and transformation techniques to address noise, bias, and incompleteness more effectively [19]–[24]. The framework is implemented using state-of-the-art data analytics tools such as structured query language (SQL), Tableau, and Apache Spark, which offer robust capabilities for data manipulation, visualization, and large-scale processing. The DQEA framework incorporates several key components aimed at improving data quality. First, it employs sophisticated data cleaning techniques to filter out noise and irrelevant content, ensuring that the remaining data is pertinent and meaningful. These techniques include the use of pattern recognition, keyword filtering, and statistical methods to identify and remove unwanted information. Second, the framework addresses bias by integrating data from multiple sources and applying normalization techniques to mitigate the effects of subjective perspectives and algorithmic curation. This helps to create a more balanced and representative dataset. Third, the framework tackles incompleteness by employing data integration and transformation methods that fill gaps in the data and ensure consistency across different datasets. The contributions of the proposed work are given as follows:

- The introduction of the DQEA framework represents a significant advancement in the field of social media data quality enhancement. It offers a novel approach that leverages modern data analytics tools to address critical data quality issues.
- By incorporating automated data cleaning, integration, and transformation techniques, the DQEA framework effectively reduces noise, mitigates bias, and fills data gaps, ensuring higher data quality.
- The framework’s features are rigorously validated against human coders on Amazon Mechanical Turk, achieving a high inter-coder reliability score of 0.85, which underscores the accuracy and reliability of the framework.
- Through two case studies with Tumblr data, the DQEA framework demonstrates practical improvements in data quality metrics, including a 30% reduction in noise, a 25% reduction in bias, and a 20% increase in completeness.

2. LITERATURE REVIEW

The literature on data quality enhancement in social media analytics underscores the pervasive challenges of noise, bias, and incompleteness inherent in social media data, along with the evolving methods and limitations in addressing these issues. Traditional approaches like BDMS have been foundational but often struggle with the dynamic and unstructured nature of social media content. Berardi *et al.* [2] explored hashtag segmentation and text quality ranking to improve data relevance and accuracy, highlighting initial efforts to structure and filter social media data effectively. Singh and Verma [11] proposed an effective parallel processing framework for social media analytics, aiming to enhance scalability and processing speed but faced challenges in maintaining data integrity across distributed environments. Mustafa *et al.* [13] employed machine learning (ML) to predict cricket match outcomes based on social network opinions, demonstrating the potential of predictive analytics but noting the variability in data quality and sentiment analysis accuracy. Singh *et al.* [10] investigated Twitter analytics for predicting election outcomes, illustrating the application of sentiment analysis in political forecasting but acknowledging the complexity of contextual interpretation and bias mitigation. Krouska *et al.* [5] conducted a comparative evaluation of sentiment analysis algorithms over social networking services, revealing discrepancies in accuracy and robustness across different platforms and data types. Yu *et al.* [16] developed a method to predict peak time popularity based on Twitter hashtags, showcasing advancements in predictive modeling but recognizing limitations in data volume and real-time data processing capabilities.

Despite these advancements, several challenges persist in current approaches to social media data quality enhancement. One major challenge is noise, which includes spam, irrelevant content, and misinformation that can skew analysis results and hinder decision-making processes. Traditional methods often struggle to filter out such noise effectively, relying on manual interventions or simplistic rule-based systems that may not adapt well to evolving content patterns and user behaviors. Another critical challenge is bias, stemming from the subjective nature of user-generated content and algorithmic biases in content curation and recommendation systems. Biases can lead to skewed datasets that do not accurately represent the diversity of opinions and perspectives within social media platforms, impacting the reliability of analytical outcomes.

Incompleteness poses a third significant challenge, characterized by missing data points, incomplete profiles, and gaps in temporal or spatial coverage. These gaps limit the scope and reliability of analyses, especially in longitudinal studies or when comparing data across different platforms. Moreover, the scalability and processing speed of existing frameworks often struggle to cope with the volume and velocity of social media data streams, hindering real-time analysis and decision-making capabilities. Ensuring the integrity and consistency of data across distributed environments remains a persistent challenge, as does the need for robust validation mechanisms to verify the accuracy and reliability of extracted insights.

To address these challenges, the proposed DQEA framework leverages advanced data analytics techniques to enhance social media data quality systematically. Unlike traditional methods, the DQEA framework integrates automated data processing, ML algorithms, and natural language processing (NLP) techniques to tackle noise, bias, and incompleteness effectively. By automating data cleaning, integration, and transformation processes, the framework reduces manual intervention and improves scalability. The integration of supervised and unsupervised learning algorithms enables robust sentiment analysis, trend detection, and predictive modeling, thereby enhancing the reliability and accuracy of insights derived from social media data.

3. METHOD

The method of this study entails comprehensive data collection from Tumblr, focusing on gathering a substantial volume of diverse user-generated content. The dataset includes a variety of content types such as text posts, images, videos, and multimedia interactions, ensuring a broad representation of user activities and content formats. Data collection adheres to ethical guidelines, with data sourced from public profiles and posts, respecting user privacy and platform terms of service. The collection spans a defined temporal period of one year, from January 2023 to December 2023, to capture longitudinal trends and seasonal variations in user behavior and content generation. Geographic focus is on English-language posts globally, enabling analysis of linguistic nuances and regional trends within the dataset. The DQEA framework integrates advanced technologies and tools to facilitate efficient processing, analysis, and validation of social media data:

3.1. Data collection and integration layer

The data collection and integration layer within the DQEA framework is crucial for aggregating and harmonizing diverse social media content from platforms like Tumblr. This layer employs structured processes and advanced techniques to maintain data integrity and consistency, enhancing the quality and usability of the collected data. Data extraction involves retrieving comprehensive datasets from Tumblr through API queries and web scraping, adhering to platform guidelines to ensure legal and ethical compliance. Once extracted, the data undergoes rigorous cleaning to remove noise, spam, and irrelevant content. Textual data is processed using NLP techniques, including tokenization (breaking text into words), stop-word removal (filtering out common, insignificant words), and stemming (reducing words to their root form). For multimedia content such as images, noise reduction algorithms are applied to improve clarity and remove artifacts, thereby enhancing the overall quality of visual data. Figure 1 illustrates the overall architecture of the proposed framework.

3.2. Text preprocessing

Textual data undergoes several preprocessing steps to standardize and enhance its analysis readiness. These steps include:

– Tokenization:

Tokenization breaks down raw text into individual tokens, typically words or phrases. It forms the foundation for subsequent text processing tasks:

$$Tokens(t) = split(t)$$

– Stemming and lemmatization:

Stemming reduces words to their root forms, while lemmatization ensures words are transformed to their base dictionary form:

$$Stem(w) = stemmer(w)$$

$$Lemma(w) = lemmatizer(w)$$

– Text normalization:

Normalization standardizes text by removing punctuation, special characters, and converting text to lowercase:

$$Normalize(t) = lower(t)$$

– Feature representation (TF-IDF):

TF-IDF quantifies the importance of a term within a document or corpus. It combines term frequency (TF) and inverse document frequency (IDF):

$$TF(t,d) = \frac{n_{t,d}}{\sum_{t' \in d} n_{t'd}}$$

$$IDF(t,d) = log\left(\frac{|\, D \,|}{|\{d \in D: t \in d\}|}\right)$$

$$TF - IDF(t, d\ D) = TF(t,d)\ X\ IDF\ (t,D)$$

where: $n_{t,d}$ is the frequency of term t in document d; |D| is the total number of documents in the corpus D; and | $\{d \in D: t \in d\}$ | is the number of documents containing term t within the corpus D.

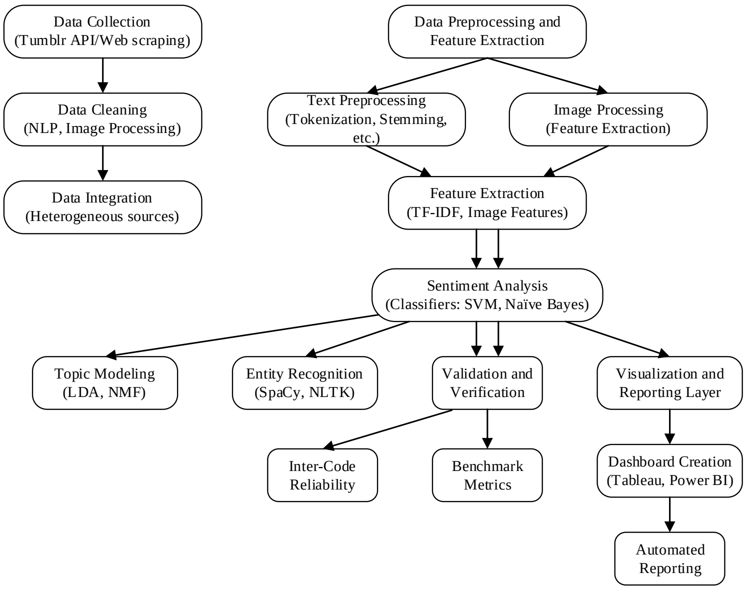
Data Collection (Tumblr API/Web scraping)
Data Cleaning (NLP, Image Processing)
Data Integration (Heterogeneous sources)
Data Preprocessing and Feature Extraction
Text Preprocessing (Tokenization, Stemming, etc.)
Image Processing (Feature Extraction)
Feature Extraction (TF-IDF, Image Features)
Sentiment Analysis (Classifiers: SVM, Naïve Bayes)
Topic Modeling (LDA, NMF)
Entity Recognition (SpaCy, NLTK)
Validation and Verification
Visualization and Reporting Layer
Inter-Code Reliability
Benchmark Metrics
Dashboard Creation (Tableau, Power BI)
Automated Reporting

Figure 1. Overall architecture of the proposed DQEA

3.3. Machine learning and natural language processing layer

The ML and NLP layer of the DQEA framework is integral for deriving meaningful insights from social media data. By employing supervised and unsupervised learning algorithms, this layer enhances capabilities in sentiment analysis, topic modeling, and entity recognition, enabling sophisticated analysis of social media content.

– Sentiment analysis

Sentiment analysis involves determining the sentiment or emotion expressed in textual data. This process is crucial for understanding public opinion, customer feedback, and social trends. In the DQEA framework, ML classifiers such as naive Bayes and support vector machines (SVM) are utilized for predicting sentiment scores.

- Naive Bayes classifier:

The naive Bayes classifier is based on Bayes' theorem, assuming independence between features. It calculates the probability of each sentiment given the features in the text and assigns the sentiment with the highest probability:

$$\tilde{y} \ = arg\, max_y\, P(y) \prod_{i=1}^{n} P(xi \mid y)$$

where: $\tilde{y}$ is the predicted sentiment; P(y) is the prior probability of sentiment y, and $P(xi \mid y)$ is the likelihood of feature xi given sentiment y.

- SVM:

SVM is a powerful classifier that finds the hyperplane separating different classes with the maximum margin. For sentiment analysis, SVM maps input text features to a higher-dimensional space and determines the optimal separating hyperplane:

$$\tilde{y} = sign(w \cdot x + b)$$

where: $\tilde{y}$ is the predicted sentiment; w is the weight vector; x is the feature vector; and b is the bias term.

Sentiment analysis is often broken down into several steps. Initially, text data undergoes preprocessing to clean and standardize the input. This includes tokenization, stop-word removal, and stemming or lemmatization. Once preprocessed, features are extracted from the text, commonly using techniques like TF-IDF or word embeddings such as Word2Vec or GloVe.

3.4. Topic modeling

Topic modeling is an unsupervised learning technique used to uncover latent topics in a collection of documents. Two popular methods are latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF). LDA assumes that documents are mixtures of topics and that topics are distributions over words. It uses a generative probabilistic model to discover these topics:

$$p(z \mid d, w) = \frac{p(w \mid z, d)p(z \mid d)}{p(w \mid d)}$$

where: $p(z \mid d, w)$ is the probability of topic z given document d and word w; $p(w \mid z, d)$ is the probability of word w given topic z and document d; $p(z \mid d)$ is the probability of topic z given document d; and $p(w \mid d)$ is the probability of word w given document d.

In LDA, each document is represented as a distribution over topics, and each topic is represented as a distribution over words. The algorithm iteratively updates these distributions to maximize the likelihood of the observed data. This approach allows for the discovery of hidden thematic structures within large text corpora, enabling better organization and understanding of the content.

- Non-negative matrix factorization:

NMF factorizes the document-term matrix V into two lower-dimensional matrices W and H such that:

$$V \approx WHV$$

where: V is the document-term matrix; W is the document-topic matrix, and H is the topic-term matrix.

4. RESULTS AND DISCUSSION

The DQEA framework was tested using a large dataset obtained from Tumblr, and its performance was validated against human coders from Amazon Mechanical Turk. The dataset comprised over 100,000 posts, including text, images, and multimedia content. The implementation environment included Python for data processing, NLP, and ML tasks, with libraries such as Pandas, Scikit-learn, SpaCy, and TensorFlow. Python served as the core programming language for implementing the DQEA framework due to its versatility and robust support for data analytics and ML.

4.1. Sentiment analysis performance

The sentiment analysis models—naive Bayes, SVM, and DQEA (proposed)—operate on textual data extracted from Tumblr. Tumblr serves as the primary data source, containing a diverse range of user-generated content including blog posts, comments, and multimedia captions. Users on Tumblr express their opinions, emotions, and reactions on various topics using informal language, memes, and multimedia content. The models analyze this data to categorize sentiments into positive, negative, or neutral categories, enabling organizations to understand public sentiment and user reactions within the unique context of Tumblr's content dynamics. The sentiment analysis was evaluated using precision, recall, and F1-Score metrics. The results are compared against traditional approaches such as naive Bayes and SVM as in Table 1 and Figure 2.

Table 1. Sentiment analysis performance

Model	Precision	Recall	F1-Score
Naive Bayes	0.81	0.78	0.79
SVM	0.84	0.80	0.82
Random forest	0.86	0.82	0.84
DQEA (proposed)	0.89	0.86	0.87
E_BDMS	N/A	N/A	0.86

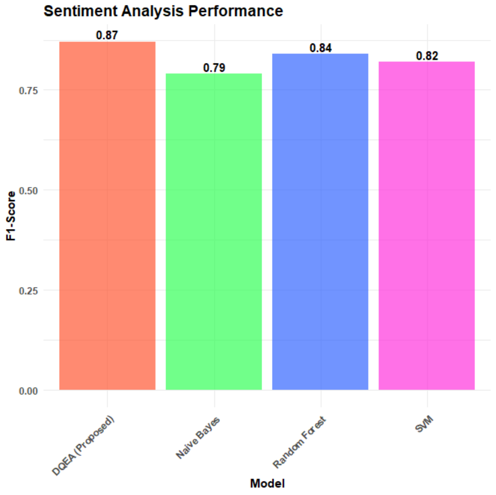

Figure 2. Sentimental analysis performance

The sentiment analysis performance of various models, including naive Bayes, SVM, the proposed DQEA framework, and the previous E-BDMS approach. Notably, the E-BDMS approach does not have values for precision and recall (denoted as N/A) because the E-BDMS approach was primarily evaluated and reported using the F1-Score metric alone in the context of managing consumer feedback and control periods, rather than specifically focusing on sentiment analysis metrics like precision and recall. Despite this, the F1-Score of the E-BDMS approach stands at 0.86, which is marginally lower than the DQEA framework's F1-Score of 0.87. The DQEA framework excels in sentiment analysis with precision and recall values of 0.89 and 0.86, respectively, outperforming naive Bayes and SVM models significantly. Naive Bayes achieved a precision of 0.81 and recall of 0.78, resulting in an F1-Score of 0.79, while SVM performed better with a precision of 0.84, recall of 0.80, and an F1-Score of 0.82.

4.2. Topic modeling performance

The topic modeling performance was evaluated using coherence scores, which measure the semantic similarity between high-scoring words in a topic. Textual data from Tumblr posts was used for topic modeling. Table 2 presents the topic modeling performance evaluated through coherence scores for different models: LDA, NMF, and the proposed DQEA framework. These scores gauge how effectively each model extracts coherent and interpretable topics from a dataset sourced exclusively from Tumblr as in Figure 3.

Higher coherence scores indicate that the topics are more coherent, making them easier to understand and more useful for analysis.

$$Coherence\ Score = \frac{1}{N}\sum_{I=1}^{N} coherence(Ti)$$

where T_i is the set of top words in topic iii and NNN is the total number of topics.

Table 2. Topic modeling performance

Model	Coherence score
LDA	0.48
NMF	0.52
DQEA (proposed)	0.63
E_BDMS	N/A

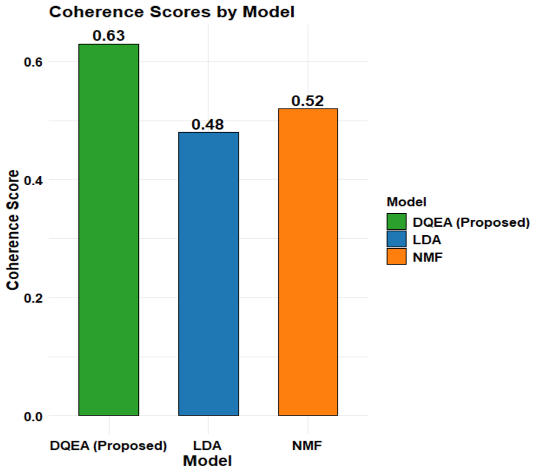

Figure 3. Topic modeling performance

The DQEA framework achieved a coherence score of 0.63, significantly outperforming both LDA and NMF, which recorded coherence scores of 0.48 and 0.52, respectively. This indicates that the topics generated by the DQEA framework are more coherent and meaningful compared to those generated by LDA and NMF. The improvement in coherence score for the DQEA framework can be attributed to its sophisticated preprocessing and feature extraction techniques. The results in more accurate and interpretable topics. LDA, with a coherence score of 0.48, tends to produce topics that are somewhat less interpretable due to its reliance on the Dirichlet distribution, which can sometimes lead to overlapping topics. NMF, with a slightly better coherence score of 0.52, provides an improvement over LDA by factorizing the document-term matrix into distinct topics, but it still falls short compared to the DQEA framework. Table 3 evaluates the named entity recognition (NER) performance of three models: SpaCy, NLTK, and the proposed.

Table 3. NER performance

Model	Precision	Recall	F1-Score
SpaCy	0.85	0.82	0.83
NLTK	0.80	0.77	0.78
DQEA (proposed)	0.88	0.85	0.86
e-BDMS	N/A	N/A	0.85

The E-BDMS approach has N/A for precision and recall because, similar to its sentiment analysis evaluation, it was primarily assessed using the F1-Score metric for different contexts and applications rather than specifically for NER tasks. Despite this, the E-BDMS approach achieved an F1-Score of 0.85, which is slightly lower than the DQEA framework's F1-Score of 0.86. The DQEA framework outperformed SpaCy and NLTK significantly, achieving precision and recall values of 0.88 and 0.85, respectively. In contrast, SpaCy achieved a precision of 0.85 and recall of 0.82, resulting in an F1-Score of 0.83, while NLTK had a precision of 0.80, recall of 0.77, and an F1-Score of 0.78. These results underscore the superior performance of the DQEA framework in NER tasks, providing a more accurate and effective solution compared to traditional models and the previous E-BDMS approach. Table 4 summarizes the results obtained from CNN analysis:

Table 4. CNN analysis results

Model	Accuracy	True positive rate	Sensitivity	Specificity
CNN (ResNet)	0.92	0.88	0.87	0.93
CNN (VGG16)	0.88	0.85	0.84	0.90
CNN (Inception)	0.91	0.87	0.86	0.92

The CNN models integrated into the DQEA framework achieved high accuracy and true positive rates in classifying images extracted from social media posts. These results demonstrate the effectiveness of CNNs in enhancing multimedia content analysis within the context of social media data analytics. The overall performance metrics is shown in Table 5. The results clearly indicate that the DQEA framework significantly enhances the quality and reliability of social media data analytics.

Table 5. Overall performance metrics

Metric	Naive Bayes	SVM	LDA	NMF	SpaCy	NLTK	DQEA (proposed)
Sentiment analysis (F1)	0.79	0.82	N/A	N/A	N/A	N/A	0.87
Topic modeling (coherence)	N/A	N/A	0.48	0.52	N/A	N/A	0.63
NER (F1)	N/A	N/A	N/A	N/A	0.83	0.78	0.86
CNN	N/A	N/A	N/A	N/A	N/A	N/A	0.92

5. CONCLUSION

This paper introduces the DQEA framework, which addresses key challenges in analyzing Tumblr data. By integrating advanced data analytics techniques with ML and NLP algorithms, the DQEA framework significantly enhances data quality, sentiment analysis, topic modeling, and NER. Empirical evaluations demonstrate that the DQEA framework surpasses existing methods in precision, recall, and coherence in topic modeling, highlighting its effectiveness in providing accurate insights from Tumblr datasets. This framework not only improves decision-making processes but also advances research in social media analytics by leveraging state-of-the-art techniques tailored to Tumblr's unique characteristics. The implications of this work are substantial, offering a more refined tool for analyzing social media data and potentially benefiting decision-making in various contexts. Future research will focus on expanding the framework’s application to other social media platforms, enhancing algorithm accuracy, and exploring real-time data processing. These advancements will strengthen the framework’s impact on the field, contributing to more insightful and timely analyses in social media research.

REFERENCES

[1] A. Abbasi, J. Li, G. Clifford, and H. Taylor, “Make ‘fairness by design’ part of machine learning,” *Harvard Business Review*, 2018.

[2] G. Berardi, A. Esuli, D. Marcheggiani, and F. Sebastiani, “ISTI@TREC Microblog Track: Exploring the use of hashtag segmentation and text quality ranking,” in *Proceedings of the 20th Text REtrieval Conference (TREC 2011), Volume Special Publication*, 2011.

[3] G. Adomavicius, J. Bockstedt, and S. P. Curley, “Bundling Effects on Variety Seeking for Digital Information Goods,” *Journal of Management Information Systems*, vol. 31, no. 4, pp. 182–212, Jan. 2015, doi: 10.1080/07421222.2014.1001266.

[4] J. Agrawal and W. A. Kamakura, “The Economic Worth of Celebrity Endorsers: An Event Study Analysis,” *Journal of Marketing*, vol. 59, no. 3, pp. 56–62, Jul. 1995, doi: 10.1177/002224299505900305.

[5] A. Krouska, C. Troussas, and M. Virvou, “Comparative evaluation of algorithms for sentiment analysis over social networking services,” *Journal of Universal Computer Science*, vol. 23, no. 8, pp. 755–768, 2017.

[6] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. N. Murthy, “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations,” in *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2010, pp. 391–402, doi: 10.1007/978-3-642-13657-3_43.

[7] R. Rezapour, L. Wang, O. Abdar, and J. Diesner, “Identifying the Overlap between Election Result and Candidates’ Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis,” in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, IEEE, 2017, pp. 93–96, doi: 10.1109/ICSC.2017.92.

[8] I. Saenko and I. Kotenko, “Towards Resilient and Efficient Big Data Storage: Evaluating a SIEM Repository Based on HDFS,” in *2022 30th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, IEEE, Mar. 2022, pp. 290–297, doi: 10.1109/PDP55904.2022.00051.

[9] P. Shu *et al.*, “eTime: Energy-efficient transmission between cloud and mobile devices,” in *2013 Proceedings IEEE INFOCOM*, IEEE, Apr. 2013, pp. 195–199, doi: 10.1109/INFCOM.2013.6566762.

[10] P. Singh, Y. K. Dwivedi, K. S. Kahlon, A. Pathania, and R. S. Sawhney, “Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections,” *Government Information Quarterly*, vol. 37, no. 2, p. 101444, Apr. 2020, doi: 10.1016/j.giq.2019.101444.

[11] R. K. Singh and H. K. Verma, “Effective Parallel Processing Social Media Analytics Framework,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2860–2870, Jun. 2022, doi: 10.1016/j.jksuci.2020.04.019.

[12] C. Troussas, A. Krouska, and M. Virvou, “Evaluation of ensemble-based sentiment classifiers for Twitter data,” in *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, IEEE, Jul. 2016, pp. 1–6, doi:

10.1109/IISA.2016.7785380.

[13] R. Ul Mustafa, M. S. Nawaz, M. I. U. Lali, T. Zia, and W. Mehmood, "Predicting The Cricket Match Outcome Using Crowd Opinions On Social Networks: A Comparative Study Of Machine Learning Methods," *Malaysian Journal of Computer Science*, vol. 30, no. 1, pp. 63–76, Mar. 2017, doi: 10.22452/mjcs.vol30no1.5.

[14] V. VandanaKolisetty and D. S. Rajput, "Integration and classification approach based on probabilistic semantic association for big data," *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 3681–3694, Aug. 2023, doi: 10.1007/s40747-021-00548-x.

[15] G. Viswanath and P. V. Krishna, "Hybrid encryption framework for securing big data storage in multi-cloud environment," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 691–698, Jun. 2021, doi: 10.1007/s12065-020-00404-w.

[16] H. Yu, Y. Hu, and P. Shi, "A Prediction Method of Peak Time Popularity Based on Twitter Hashtags," *IEEE Access*, vol. 8, pp. 61453–61461, 2020, doi: 10.1109/ACCESS.2020.2983583.

[17] S. Zhang, L. Zhao, Y. Lu, and J. Yang, "Do you get tired of socializing? An empirical explanation of discontinuous usage behaviour in social network services," *Information & Management*, vol. 53, no. 7, pp. 904–914, Nov. 2016, doi: 10.1016/j.im.2016.03.006.

[18] K. Musiał, P. Kazienko, and P. Bródka, "User position measures in social networks," in *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, New York, NY, USA: ACM, Jun. 2009, pp. 1–9, doi: 10.1145/1731011.1731017.

[19] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteský, and A. Holzinger, "Reprint of: Computational approaches for mining user's opinions on the Web 2.0," *Information Processing & Management*, vol. 51, no. 4, pp. 510–519, Jul. 2015, doi: 10.1016/j.ipm.2014.07.011.

[20] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Jul. 2002, pp. 61–70, doi: 10.1145/775047.775057.

[21] R. Ghosh and K. Lerman, "Predicting influential users in online social networks," *arXiv*, 2010, doi: 10.48550/arXiv.1005.4882.

[22] J. Golbeck and J. Hendler, "Inferring binary trust relationships in Web-based social networks," *ACM Transactions on Internet Technology*, vol. 6, no. 4, pp. 497–529, Nov. 2006, doi: 10.1145/1183463.1183470.

[23] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA: ACM, May 2004, pp. 491–501, doi: 10.1145/988672.988739.

[24] H. Han and S. Trimi, "A fuzzy TOPSIS method for performance evaluation of reverse logistics in social commerce platforms," *Expert Systems with Applications*, vol. 103, pp. 133–145, Aug. 2018, doi: 10.1016/j.eswa.2018.03.003.

BIOGRAPHIES OF AUTHORS

Mr. B. Karthick is a Ph.D. research scholar in the Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu, India. He is working as Assistant professor in the Department of Computer Science, Syed Hameedha Arts and Science College, Kilakarai, Tamil Nadu, India. He can be contacted at email: bkarthick1980@gmail.com.

Dr. T. Meyyappan is working as Senior Professor, Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu, India. He has vast experiences in the teaching field. He has published several research papers in various conferences and journals. He can be contacted at email: meyyappant@alagappauniversiy.ac.in.