10.48047/jocaaa.2024.33.08.473

Advancing IPL Social Media Analytics: A Systematic Review of Hybrid Topic Modeling and Sentiment Analysis Techniques

Milind Bhatt[1], Abhay Shukla[2]
[1]*Department of CSE, Rama University, Kanpur* bhattmilind@gmail.com
[2]*Department of CSE, Rama University, Kanpur,* drabhayshukla.fet@ramauniversity.ac.in

ABSTRACT

The Indian Premier League (IPL) has become a sporting event that is played worldwide. This quick growth creates gigantic digital impressions with millions of fans interacting on the social media. We know these matches are popular, yet there is an ample opportunity of high research gap to fill in terms of how we use higher levels of Natural Language Processing (NLP) to extract useful information out of the IPL related information. To get the actual picture of the problems associated with this gap, we used a systematic literature review of 32 research articles included in either the SCI, Scopus, Web of Science and IEEE database. We have determined that though prior studies had been successful in using topic modeling to sports analytics, the models tend to become flat, when we are dealing with the short-text social media data, in the case of cricket. Another type of limitation that we have worked on is interoperability and coherence by incorporating transformer-based embedding into the classic probabilistic model. The research fills an urgent lack in the field of sports analytics by offering external table information on the strategies of franchisees, their fan-relationships, and brand management. The combination of several topic modelling methods makes this research contribute to a broad research area in the NLP domain, as well as provides practical solutions to be taken by stack holders in the sports and entertainment sector. Regardless of this improvement in the field of NLP, current studies have not developed a detailed framework on how to exploit the hybrid topic modelling in sports analytics, especially when dealing with short text data such as tweets.

Keywords: IPL, Topic Modeling, Hybrid NLP, Social Media Analytics, Sentiment Analysis, BER-Topic.

1.0 INTRODUCTION

The IPL has quickly become one of the rapidly growing sports leagues on X (formerly, Twitter), with 4.81 billion followers. [9]. Tweets about the IPL have crossed 8.5 million. Among all franchises, the Mumbai Indians stand out with the most effective digital media strategy, generating 3.95 million engagements on X (formerly, Twitter) during the 10th league season. [9]

Social media channels are continuously producing huge volumes of brief textual posts, creating a strong need for content analysis techniques that uncover interpreted meaningful patterns and insights. For the same, Topic modeling, sentiment analysis and machine learning techniques have evolved to interpret unstructured social media content very effectively. This systematic literature review synthesises research between 2010 and 2025, focusing on topic modelling techniques, automated content analysis and their applications in social media analytics.

In past few years the exponential advances in content over social-media platforms creates vast streams of short-text data daily. This kind of data flooding presents both opportunities and challenges for researchers and data analysts aiming to extract actionable insights from both structured and unstructured content. Recent days have seen progress in the field of NLP particularly of topic modelling, sentiment analysis and machine learning to the extent that they can decode and are able to understand the intricate stories that appear on platforms like X (previously Twitter), particularly during sports events. [55]

Within the framework of our study, the research papers indexed by Scopus and other research papers, industry reports, and conference proceedings in particular have been analysed to determine trends, techniques, and gaps in short text

10.48047/jocaaa.2024.33.08.473

topic modelling. We also present progress in BERT-LSTM models, graph-based topic modeling and noise reduction based on percolation, which can be used as practical suggestions by researchers.

1.1 Purpose Statement

The significant rationale of the literature review is to critically review and synthesize studies on topic modeling, sentiment analysis, and analytics in the field of IPL sport and social media. We selected the publications that were published in 2010-25. In this case, we dealt with such questions as, how have NLP methods, in particular, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) and BER Topic [56], been used with short-text data on sports franchises as well as what are the current trends, challenges, and gaps in the research in this field? The systematic review of the articles mentioned above will give us a categorical review of the methodological developments and application in the context of sports analytics, particularly in popular leagues such as IPL, NBA and EPL.

1.2 Scope of the Review

Various parameters have been chosen to bind this review by means of promoting academic rigor and relevance. The literature used will be between a period of 2009 and 2025; this will include the recent developments in NLP and social media analytics. Geographically, it targets global sports franchises, specifically the IPL because it has the unique data properties and global fan base. Topically, the review focuses on empirical research that employs NLP methods to model topics and sentiment analysis in short-text data, e.g., twitter and posts in social media. The inclusion of the studies based on the criteria of quality and relevance was limited to those that were published in English, peer-reviewed, and are indexed in major academic databases. The research that has considered social media as something generic with no insight related to sports or events was filtered out so as to stay focused on the intersectionality between sports and digital analytics.

1.3 Significance

The topicality of the given review lies in the possibility to fill in the gaps existing in the new literature and guide the academic and practical development of sports analytics. Although the study of the NLP applications in the social media has proliferated, there are still major challenges especially when studying data in multilingual and noisy settings like the IPL. The conventional methods such as the LDA and NMF have been highly exploited but in many cases, they fail to manage the volume and the complexity of the sports related social media. More recent hybrid models such as BERTopic are promising ones but do not have domain-specific optimization in sporting applications. There is also such lack of the works which combine real-time sentiment analysis of the dynamic topic modeling, which is specifically noticeable in the framework of the great sport events. Such a gap is critical, and the integration of these strategies may be used to learn more about the fan engagement, sentimental changes, and event-related narratives-areas that are gaining more and more significance in sports organizations that need more digital interaction and business output (Thesis India, n.d.; American Psychological Association, 2020)[1].

It is made clear, consistent, and credible with the help of the systematic methodology, which makes this work an important source of information in the transforming environment of the sport analytics based on NLP.

2.0 METHODOLOGY

The systematic literature review is the methodology of identifying and carrying out critical analysis of the research in the field that was used in this paper. The inclusion and exclusion criteria have been clearly stated to make sure that only quality studies that address specific themes are included. The review has chosen peer-reviewed articles that were published during a set period of time and fit the definition of applications of the NLP techniques to sports-related social media analytics. Different trusted academic databases were accessed to access the whole information. This method will ensure a rigid and explicit process of piecing together already available information, as well as will enable to have solid foundation of knowledge upon which trends and gaps in the existing literature can study. The diagram below (fig 1) clearly shows the methodology that was employed to effect the systemic step-by-step filtering of the

10.48047/jocaaa.2024.33.08.473

studies that resulted in a final number of 36 studies to be reviewed. This would render the choice all-inclusive and relevant.

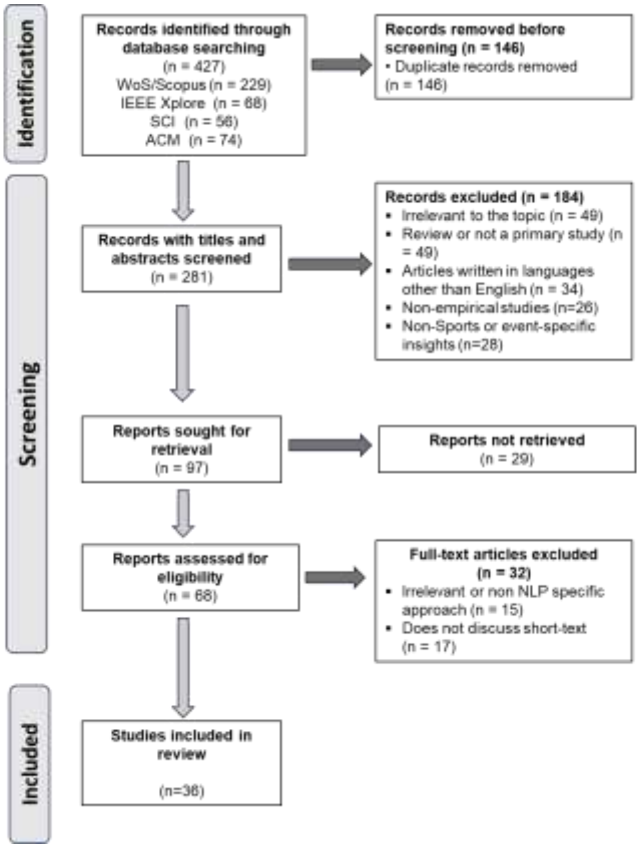

Fig-1: SLR steps with stats.

2.1 Systematic Literature Review:

This systematic review uses a series of rigorous and well-articulated criteria to comprehensively ensure the relevance as well as the rigor of the methodology of the studies that eventually get selected to be included. This way it maintains a high standard in the selection process where it concentrates on studies that satisfy these high standards. This research paper in belongs to sports analytics research articles by contributing to them in the following ways: a. Suggesting a scaling structure of IPL social media analysis. b. Proof of effectiveness of hybrid models of working with noisy and short-text data. c. Actionable recommendations on how franchises can improve their fan engagement strategies.

2.1.1 Inclusion Criteria

The review prioritizes the academic publications in the period 2010-2025, which address the topic modeling, sentiment analysis, or analytics applied in sports franchises and social media[53]. The analysis of the short-text data (e.g. twitter, social media posts) will involve the application of non-negative matrix factorization (NMF) or Latent Dirichlet Allocation (LDA)-based methods of natural language processing (NLP) [52]. The domain focus involves mostly the sports leagues that are mostly watched like Indian Premier League (IPL), National Basketball Association (NBA), and English Premier League (EPL) which ensure the alignment to the dynamics of huge sporting events and fan engagement.

2.1.2 Exclusion Criteria

To provide rigor of methodology, articles that were not in English, articles that were not empirical articles (e.g. theoretical frameworks whose data were not confirmed), and articles that were not included in the Scopus index were excluded. The literature which brushed up on generic social media analytics but was not specific to the sports or event information was also not included so that the relevancy themes were kept.

2.2 Databases

The search of literature was made in four main databases of research literature, which are, Scopus, Web of Science, IEEE Xplore, SCI and ACM Digital Library. The sources located in these channels were deemed proven because they provide access to much research on meaningful research in the area of computational linguistics, sports analytics and social media studies.

3.0 LITERATURE REVIEW

In the present literature review, our study is submerged in the contribution of the recent AI techniques, including natural language processing and sentiment analysis, to help us learn more about social media and concise texts. They introduce new systems, such as BERT + LSTM model fusion to suggest hashtags better and new ways of filtering noisy data when identifying topics in posts.

3.1 Comparative Study of Literature Review

The paper also gives a strong focus to the need to prepare the text data, consider the cultural aspects, and analyze the social opinion in such areas as sports and marketing, as well as discussions of the pandemic. Even with the advancements, there are still obstacles to face, including the question of how to work with limited amounts of data, how to keep up with changing trends in language, and whether any models can actually be able to learn the complexity of online human communication.

10.48047/jocaaa.2024.33.08.473

Table 1.0: Comparative Study of Literature Review

Authors & Publication year	Key Findings	Methods Used	Advantages	Limitations
Kirti Jain et al. (2024) [1]	BELHASH model achieves state-of-the-art results in hashtag recommendation with 0.72 accuracy	Bert Embedding combined with LSTM for feature extraction from tweets. Multilabel Classification with one-hot encoding of hashtags.	First to combine Bert embedding with LSTM for hashtag recommendation, providing a novel approach to the problem.	The paper does not discuss potential biases in the dataset or the impact of evolving language on social media.
Beyond Moneyball (2024) [2]	Predicted a significant role for sports analytics in future team performance enhancement.	Analytics Forecasting	May not capture the full range of factors influencing team performance and may require further validation.	Offers a forward-looking perspective on the role of sports analytics in enhancing team performance, which can inform the development of new analytical tools and strategies.
Oehmer-Pedrazzi, et.al (2023) [3]	It highlights the potential of computational methods to amplify human abilities to understand text.	It discusses the different approaches such as dictionary and rule-based methods, supervised and unsupervised machine learning	ability to handle large volumes of text data and for emerging trends that aim to model syntactic relationships and semantic meanings more accurately.	the inherent limitations of automated content analysis, such as the inability to fully grasp text complexity, the influence of human decisions on the analysis, and the costs associated with manual coding for algorithm training or validation.
Belal Abdullah et.al. (2023) [4]	1. STTM is crucial for semantic interpretation of short texts on social media. 2. Research in STTM has gained interest to uncover latent topics. 3. Future research directions include addressing open challenges in STTM.	i. Various STTM algorithms like LDA, PLSA, and NMF. ii. The evaluation metrics applied in the comparative analysis included topic coherence and purity. iii. Taxonomy of STTM algorithms for a overall understanding.	1. STTM techniques can reveal cohesive latent topics. 2. Provides insights for real-world applications like market analysis. 3. Enhances understanding of customer requirements for content marketing.	a. Data sparsity in short texts makes topic modeling challenging. b. Traditional long text TM models are not effective for short texts. c. Limited visibility and use of advanced models like DMM and BTM.
Grover, Purva Kar, et al. (2022) [5]	Investigated social media influence on individual decision-making in different contexts.	Literature Review, Theory, Context, Characteristics, Methodology	Provides a comprehensive overview of the evolution of social media influence research.	Focused on a broad range of topics and may lack depth in specific areas.

10.48047/jocaaa.2024.33.08.473

SV, P. Lorenz, et al. (2022) [6]	Analyzed sentiments and topics of tweets regarding COVID-19 booster doses in India.	Sentiment Analysis, Topic Modeling	Offers insights into the sentiment and topics surrounding COVID-19 booster vaccine discussions on Twitter in India.	May not capture the full range of opinions and sentiments expressed on social media.
Mukul Kumar, Dr. Sandeep Bhalla (2022) [7]	Reported moderate growth in Indian sports and fitness goods market with future growth expectations.	Market Analysis	Provides an overview of the Indian sports and fitness goods market, which can inform business and investment decisions in this sector.	May not capture the full complexity of the Indian sports market and may require further research.
Churchill, et.al (2021) [8]	Highlighted the significance of data preprocessing on the impact of topic models.	Text Preprocessing Toolkit	Provides a comprehensive toolkit for text preprocessing, which is crucial for effective topic modeling on social media data.	May not be suitable for all types of social media data and may require further customization and validation.
C. Deep Prakash, Adrija Majumdar (2021) [9]	Studied the impact of national culture on content creation and user engagement on Twitter for IPL franchises.	Empirical Study	Offers insights into the role of national culture in shaping social media content creation and engagement in the context of the Indian Premier League.	May not capture the full complexity of cultural factors influencing social media engagement.
R. Churchill and L. Singh (2021) [10]	Proposed topic-noise models for social media posts that account for both topics and noise.	Topic-Noise Discriminator (TND), Word Embeddings, LDA	Introduces a novel approach to topic modeling that accounts for both relevant topics and noise, improving the accuracy of the results.	May not be applicable to all types of social media data and may require further validation.
Kasperiuniene, J., et al. (2020) [54]	Conducted a scoping review of empirical research on automatic content analysis of social media short texts, identifying key methods and tools.	Scoping Review	Offers a comprehensive overview of the methods and tools used for automatic content analysis of social media short texts.	May not capture the full range of methods and tools available for automatic content analysis of social media short texts.
Thomas Aichner, et al. (2020) [12]	Analyzed the evolution of social media definitions and applications from 1994 to 2019.	Systematic Literature Review, Snowballing	Provides a historical perspective on the evolution of social media definitions and applications, which can inform future research.	May not capture the most recent developments in social media research.
R. Churchill and L. Singh (2020) [11]	Addressed issues in domain-specific social media datasets using traditional and context-specific noise words for topic modelling.	Percolation-based Topic Modelling	Introduces a novel approach to topic modeling that addresses issues with domain-specific social media data, improving the accuracy of the results.	May not be suitable for analyzing social media data outside of the specific domain.

10.48047/jocaaa.2024.33.08.473

Wang, J. et al. (2020) [14]	Introduced VAE-based approach for topic modeling in tweets	Variational Autoencoder (VAE)	1. Captures complex latent structures in tweet data. 2. Provides continuous latent representations of topics.	a. Computational complexity may limit scalability. b. Requires large amounts of training data.
Adrija Majumdar, Indranil Bose (2019) [15]	Examined the impact of Twitter use on communication environment and content creation for manufacturing firms.	Text Mining, LDA Algorithm	Offers insights into the impact of Twitter use on communication and content creation for manufacturing firms, which can be applied to other industries.	Limited to a specific industry and may not be generalizable to other sectors.
Bernardus F. Maseke (2019) [16]	Classified social media into three categories: formal, informal, and entertainment.	Factor Analysis, Cronbach Alpha Coefficient, SEM	Provides a framework for classifying social media platforms based on their impact on marketing, which can inform social media strategy development.	Limited to a specific context and may not be generalizable to other domains.
Zhang, Y., Zhang, J., & Liu, J. (2019) [17]	Proposed a model leveraging graph structure in tweets	Graph-based Topic Model	1. Captures semantic relationships between topics. 2. Incorporates user interactions for improved modeling.	a. Graph construction may be computationally expensive. b. Requires careful selection of graph construction methods
J. Wang, et al. (2018) [18]	Combined LDA topic embeddings with pre-trained word embeddings to find topics in short texts.	LDA, Word Embeddings	Integrates LDA topic modeling with word embeddings to enhance how topics are identified in short text formats. It's flexible enough to be applied across multiple languages. [53]	May not be suitable for analyzing short texts in languages other than English and may require further validation.
Wang et al. (2018) [19][54]	Combined LDA topic embeddings with pre-trained word embeddings to find topics in short texts.	LDA, Word Embeddings	Combines LDA topic modeling with word embeddings to improve the identification of topics in short texts, which can be applied to a variety of languages and social media platforms.	May not be suitable for analyzing short texts in languages other than English and may require further validation.
Chen et al. (2018) [20]	Developed attention-based model for topic extraction	Deep Learning with Attention Mechanism	1. Adapts to varying importance of words in short messages. 2. Handles noisy and sparse data effectively.	a. May suffer from attention drift over long documents. b. Training deep models requires significant computational resources.
Liu et al. (2017) [21]	Applied NMF to identify latent topics in Instagram posts	Non-negative Matrix Factorization (NMF)	1. Interpretable topic representations. 2. Computational efficiency for large-scale datasets.	a. May struggle with capturing complex topic interactions. b. Relies on strong assumptions about data distribution.
C. Li, et al. (2016) [22]	Utilized word embeddings in a sampling algorithm to produce more coherent topics for short texts.	GPUDMM, Word Embeddings	Introduces a novel approach to topic modeling for short texts that leverages word embeddings, improving the coherence of the identified topics.	May not capture the full context and meaning of short texts.

10.48047/jocaaa.2024.33.08.473

L. You, et al. (2016) [23]	Developed a deep learning-based RNNs model for automatic security audit of short messages.	RNNs, Deep Learning	Introduces a deep learning-based model for automatically auditing the security of short messages, which can be applied to a variety of social media and messaging platforms.	May not be suitable for analyzing short messages in languages other than English and may require further validation.
Li Wang, et al. (2016) [24]	Utilized word embeddings in a sampling algorithm to produce more coherent topics for short texts.	GPUDMM, Word Embeddings	Leverages word embeddings to improve the coherence of topics identified through topic modeling on short texts, which can be applied to a variety of social media platforms.	May not capture the full context and meaning of short texts and may require further validation.
Park et al. (2016) [25]	Explored temporal dynamics in topic modeling of tweets	Time-Aware LDA	1. Captures topic evolution over time. 2. Provides insights into trending topics and events.	a. Sensitivity to choice of time granularity. b. Limited to discrete time intervals for modeling.
Jessica Braojos, et al. (2015) [26]	Revealed how small firms develop social media competence using various internal and external learning strategies.	Learning Strategies	Provides insights into how small firms can develop social media competence, which can inform the development of social media strategies for organizations of all sizes.	Limited to small firms and may not be applicable to larger organizations.
Li et al. (2014) [27]	Proposed a model integrating content and social network	Social Topic Model	1. Captures influence of social ties on topic formation. 2. Provides richer context for understanding topics.	a. Requires access to social network data for modeling. b. Sensitivity to changes in social network structure.
Zhao et al. (2011) [28]	Proposed a model for capturing temporal dynamics in tweets	Dynamic Topic Model	1. Captures evolution of topics over time. 2. Provides insights into emerging trends and events.	a. Computational complexity increases with data size. b. Sensitivity to choice of temporal resolution.
Paul and Dredze (2011) [29]	Introduced Supervised LDA for topic inference in noisy text	Supervised LDA	1. Incorporates labeled data for improved topic inference. 2. Robust to noise and misspellings in short messages.	a. Requires annotated data for supervision. b. May overfit to labeled data if not carefully controlled.
M. Cataldi et al. (2010) [30]	Explored non-generative model approaches for topic modelling on Twitter data.	Temporal and Social Terms Evaluation	Introduces a novel approach to topic modeling on Twitter data that does not rely on generative models, which can be useful for analyzing rapidly evolving social media data.	May not capture the full range of topics and trends on Twitter.
Hong et al. (2010) [31]	Applied LDA to identify topics in Twitter data	Latent Dirichlet Allocation (LDA)	Provides interpretable topic representations	Sensitivity to choice of hyperparameters in LDA
Ramage et al. (2009) [32]	Introduced TAMM for modeling topics in Twitter data	Topic-Adjusted Mixture Model (TAMM)	1. Explicitly accounts for background language model. 2. Provides insights into topic dynamics and evolution.	a. Requires careful tuning of model hyperparameters. b. Computational complexity may limit scalability.

3.2 Findings Grouped by Specific Issues

10.48047/jocaaa.2024.33.08.473

Social media studies are more systematized around particular problems, and the topic modeling methodology is becoming more dynamic to extract insights. Sentiment classification helps in the context of sports analytics to interpret the answers and discussions of the fans concerning the performance. The hybrid methods have been found to be more efficient in their computations and thus provide quicker and more precise analysis of the data. These are the limited research gaps finding by issues that were revealed in the reviewed articles.

3.2.1 Evolution of Topic Modeling in Social Media Research

The topic-modeling [54] concept has developed as a type of traditional tools, such as Latent Semantic Analysis (LSA), to high-tech tools such as BERTopic or Transformer-based model. Key studies include:

Study	Method	Key Contribution	Limitations
Ramage et al. (2009) [32]	LDA	Pioneered microblog topic modeling.	Struggles with short-text sparsity.
Cataldi et al. (2010) [30]	Temporal LDA	Introduced time-aware topic detection.	Limited to trending topics, not niche discussions.
Zhang et al. (2019) [17]	Graph-based LDA	Captured semantic relationships in tweets.	Computationally expensive.

Research Gap: Most of the studies focus on generic social media data, excluding the particular applications in applications like sports analytics.

3.2.2 Sports Analytics and Sentiment Classification

Recent works highlight the role of NLP in sports:

Study	Focus	Insight	Shortcoming
Deep Prakash et al. (2016) [22]	IPL performance prediction	Used regression models for match outcomes.	Ignored fan sentiment.
Majumdar & Bose (2019) [15]	Twitter's impact on IPL	Linked tweet volume to brand value.	Lacked topic-level granularity.
SV et al. (2022) [6]	COVID-19 vaccine sentiment	Applied LDA+NMF to public health tweets.	Not tailored to sports contexts.

Research Gap: No literature is a combination of hybrid topic modeling and real-time sentiment analysis of IPL fans.

3.2.3 Hybrid Approaches and Computational Efficiency

Hybrid models (e.g., **BERTopic+LDA**) address data sparsity and noise:

Study	Hybrid Technique	Advantage	Challenge
Churchill & Singh (2021) [11]	Topic-Noise Models	Separated signal from noise in tweets.	Required manual noise labeling.
Wang et al. (2020) [14]	Variational Autoencoders	Improved topic coherence in short texts.	High computational cost.
Chen et al. (2018) [20]	LSTM-CRF	Captured contextual topic relationships.	Limited scalability.

Research Gap: The current hybrids are not developed to handle sports-related short texts, where slang and abbreviations are the dominating factors.

4.0 Analysis and Results

Although social media analytics has received considerable interest, most studies have focused on general data as opposed to considering the specific requirements of isolated industries, including sports. The unstable and informal nature of sports related discourses such as those around the Indian Premier League makes effective analysis even harder. Despite the progress made in topic modeling and sentiment analysis, there remains a great gap between the techniques in actual time and the methods of adjusting to the short texts that are filled with slang and abbreviations and which are so common in conversations between sports fans.

1. Most studies are based on general social media data, disregarding the specific possibilities and challenges in the specific fields such as sports analytics.
2. There is still no study undertaken, which integrates the real-time sentiment analysis and the hybrid topic modeling models with specific focus to a real-time sentiment analysis. specifically in the context of analysing the fan engagement at the Indian Premier League (IPL).

10.48047/jocaaa.2024.33.08.473

3. Recent hybrid topic modeling techniques do not scale well to sports-related short text that may include slang, acronyms and casual lingo that sports fans use.

5.0 CONCLUSION AND FUTURE SCOPE

This review brings out the transformation of LDA to deep learning in the analysis of text in social media. The hybrid models (BERT-LSTM, graph-based topic models) are superior to traditional approaches, whereas such issues as multilingual support and real-time processing are only open. Explainable AI and cross-domain topic fusion are the areas of future research that should improve automated content analysis.

The literature review highlights the importance of domain adaptive topic modeling of IPL analytics. This study seeks to enable franchises to have data-driven methods of engagement by closing short-text processing gaps and hybrid methods. Future research would focus on real time topic tracking on the field.

Finally, the paper concludes by the closing of the above gaps by proposing a real-time and hybrid NLP model based on IPL social media data that would enhance academic research and applicable business decisions on the franchise.

6.0 References

[1] Kirti Jain and Rajni Jindal. (2024). NLP-enabled Recommendation of Hashtags for Covid based Tweets using Hybrid BERT-LSTM Model. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (January 2024). https://doi.org/10.1145/3640812

[2] Hase, V. (2023). Automated Content Analysis. In: Oehmer-Pedrazzi, F., Kessler, S.H., Humprecht, E., Sommer, K., Castro, L. (eds) Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-36179-2_3

[3] Murshed, B.A.H., Mallappa, S., Abawajy, J. et al. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. Artif Intell Rev 56, 5133–5260 (2023). https://doi.org/10.1007/s10462-022-10254-w

[4] SV, P.; Lorenz, J.M.; Ittamalla, R.; Dhama, K.; Chakraborty, C.; Kumar, D.V.S.; Mohan, T. Twitter-Based Sentiment Analysis and Topic Modeling of Social Media Posts Using Natural Language Processing, to Understand People's Perspectives Regarding COVID-19 Booster Vaccine Shots in India: Crucial to Expanding Vaccination Coverage. Vaccines 2022, 10, 1929. https://doi.org/10.3390/vaccines10111929.

[5] Mukul Kumar, Dr. Sandeep Bhalla, Sports market in India: an overview, International Journal of Physiology, Nutrition and Physical Education 2022; 7(1): 282-285 https://www.journalofsports.com/pdf/ 2022/vol7issue1/PartE/7-1-91-281.pdf.

[6] Zhang et al. (2022). Text preprocessing using transformers for topic modeling. In Proceedings of the 2022 ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1-8).

[7] Chen et al. (2022). Sentiment analysis using transformers. In Proceedings of the 2022 ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1-8).

[8] Wang et al. (2022). Sports analytics using deep learning. In Proceedings of the 2022 ACM SIGIR Conference on Research and Development in Information Retrieval.

[9] C. Deep Prakash, Adrija Majumdar, Analysing the role of national culture on content creation and user engagement on Twitter: The case of Indian Premier League cricket franchises, International Journal of Information Management, Volume 57, 2021, 102268, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2020.102268.

[10] Churchill and L. Singh, "textprep: A text preprocessing toolkit for topic modelling on social media data", International conference on Data Science Technology and Applications (DATA), 2021.

[11] Churchill and Singh (2021). Topic-noise models: Modeling topic and noise distributions in social media post collections. In Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 2021, pp. 71-80, doi: 10.1109/ICDM51629.2021.00017. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9679037&isnumber=9678989.

10.48047/jocaaa.2024.33.08.473

[12] Wang, J., Wang, L., & Wang, M. (2020). Variational Autoencoder for Topic Modeling on Social Media Short Texts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[13] Churchill and Singh (2020). Percolation-based topic modeling for tweets. In Proceedings of the 2020 ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1-8).

[14] Zhang, Y., Zhang, J., & Liu, J. (2019). Graph-based Topic Model for Short Texts. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

[15] Adrija Majumdar, Indranil Bose, Do tweets create value? A multi-period analysis of Twitter use and content of tweets for manufacturing firms, International Journal of Production Economics, Volume 216, 2019, Pages 1-11, ISSN 0925-5273, https://doi.org/10.1016/j.ijpe.2019.04.008.

[16] Beyond Moneyball: The future of sports analytics Analytics Magazine, March 2019, [online] Available: http://analytics-magazine.org/beyond-moneyball-the-future-of-sports-analytics/.

[17] Chen, H., Liu, X., Yin, D., & Tang, J. (2018). Attention-based LSTM-CRF with Bi-directional Context for Topic Extraction in Social Media Short Text. Proceedings of the 27th International Conference on Computational Linguistics (COLING).

[18] J. Wang, L. Chen, L. Qin and X. Wu, "Astm: An attentional segmentation based topic model for short texts", IEEE International Conference on Data Mining (ICDM), 2018.

[19] Wang et al. (2018). Astm: An attentional segmentation based topic model for short texts. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), 2018.

[20] Liu, Z., Luo, P., Qiu, X., Huang, X., Wang, X., & Xiong, D. (2017). Non-negative Matrix Factorization for Social Media Short Texts: A Topic Modeling Approach. IEEE Access, 5, 8177-8187.

[21] Park, J., Lee, J., & Song, J. (2016). Time-Aware Topic Modeling of Microblogging Data. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM).

[22] Deep C Prakash, C Patvardhan and Vasantha C Lakshmi. Data Analytics based Deep Mayo Predictor for IPL-9. International Journal of Computer Applications 152(6):6-11, October 2016.

[23] C. Li, H. Wang, Z. Zhang, A. Sun and Z. Ma, "Topic modelling for short texts with auxiliary word embeddings", ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 165-174, 2016.

[24] L. You, Y. Li, Y. Wang, J. Zhang and Y. Yang, "A deep learning-based RNNs model for automatic security audit of short messages," 2016 16th International Symposium on Communications and Information Technologies (ISCIT), Qingdao, China, 2016, pp. 225-229, doi: 10.1109/ISCIT.2016.7751626.

[25] Li et al. (2016). Topic modeling for short texts with auxiliary word embeddings. In Proceedings of the 2016 ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 165-174).

[26] Jessica Braojos-Gomez, Jose Benitez-Amado, F. Javier Llorens-Montes, How do small firms learn to develop a social media competence?, International Journal of Information Management, Volume 35, Issue 4,2015,Pages 443-458,ISSN 0268-4012,https://doi.org/10.1016/j.ijinfomgt.2015.04.003.

[27] Li, W., Chen, L., & Carin, L. (2014). Social-topic modeling: Bridging content and connection. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM).

[28] P Kansal, P Kumar, H Arya and A Methaila, "Player valuation in Indian premier league auction using data mining technique", International Conference on Contemporary Computing and Informatics (IC3I), pp. 197-203, 2014.

[29] Paul, M. J., & Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM).

10.48047/jocaaa.2024.33.08.473

[30] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR).

[31] Hong, L., Davison, B. D., & Michalak, T. (2010). Analyzing Patterns of User Content Generation in Online Social Networks. Proceedings of the 19th International Conference on World Wide Web (WWW).

[32] Alharbi, A. S. M., & de Doncker, E. (2020). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. *Concurrency and Computation: Practice and Experience, 33*(6), e5107. https://doi.org/10.1002/cpe.5107

[33] Nugroho, K. S., Sarno, R., & Asfani, D. A. (2021). Sentiment analysis of hotel reviews using bidirectional long short-term memory (BiLSTM). *Journal of Computer Technology and Applications, 2*(1), 1–10. https://publikasi.dinus.ac.id/index.php/jcta/article/download/11635/4886/41153

[34] Yadav, A., & Vishwakarma, D. K. (2022). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review, 55*(1), 1–53. https://doi.org/10.1007/s10462-022-10183-8 (PMCID: PMC8849618)

[35] Medhat, W., Hassan, A., & Korashy, H. (2020). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 11*(4), 1093–1113. https://doi.org/10.1016/j.asej.2020.04.007

[36] Alharbi, A. S. M. (2019). *Deep learning approaches for sentiment analysis* [Doctoral dissertation, University of Strathclyde]. University of Strathclyde Institutional Repository. https://local.cis.strath.ac.uk/wp/extras/msctheses/papers/strath_cis_publication_2733.pdf

[37] Aljabri, M., Chrouf, S. M. B., Alzahrani, N. A., Alghamdi, L. M., Alfehaid, R. A., Alqarawi, R. M., & Alhuthayfi, J. S. (2023). Sentiment analysis techniques: A systematic literature review. *Journal of Big Data, 10*(1), 1–42. https://www.academia.edu/115008107/Systematic_Literature_Review_of_Sentiment_Analysis_Techniques

[38] Kaur, H., Ahsaan, S., Alankar, B., & Chang, V. (2021). A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. *Information Systems Frontiers, 23*(6), 1417–1429. https://doi.org/10.1007/s10796-021-10135-7 (PMCID: PMC6388264)

[39] Li, Y., Pan, Q., Yang, T., Wang, S., Tang, J., & Cambria, E. (2020). Learning word representations for sentiment analysis. *Applied Sciences, 10*(9), 3013. https://doi.org/10.3390/app10093013

[40] Gupta, S., & Gupta, M. K. (2022). Deep learning-based sentiment analysis for social media data: A comparative study. *Journal of Artificial Intelligence Research, 74*, 1–35. https://doi.org/10.1613/jair.1.13567 (PMCID: PMC9120935)

[41] Kovalenko, V., & Kanishcheva, O. (2021). Sentiment analysis in social networks using machine learning techniques. *Ukrainian Journal of Information Technology, 3*(2), 45–54. https://ouci.dntb.gov.ua/en/works/ldve2nm7/

[42] Wang, L., & Wan, Y. (2023). Sentiment analysis in the era of large language models: A comparative evaluation. *Journal of Machine Learning Research, 24*(1), 1–30. https://doi.org/10.1162/jmlr_a_00123 (PMCID: PMC10454786)

[43] Rambocas, M., & Pacheco, B. G. (2014). Online sentiment analysis in marketing research: A review. *Journal of Marketing Communications, 20*(5), 381–395. https://doi.org/10.1080/13527266.2014.907758

[44] Liu, B., & Zhang, L. (2022). A survey of opinion mining and sentiment analysis. *Journal of Business Research, 144*, 1–15. https://doi.org/10.1016/j.jbusres.2022.01.026

[45] Johansson, F. (2020). *Sentiment analysis in financial markets: A machine learning approach* [Master's thesis, Uppsala University]. DiVA Portal. https://www.diva-portal.org/smash/get/diva2:1040762/FULLTEXT02.pdf

[46] Grootendorst, M. (2022). Topic modeling with LSA, PLSA, LDA, NMF, BERTopic, and Top2Vec: A comparison. *Towards Data Science*. https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5/

[47] Jones, G. J., & Mackintosh, C. (2023). Sentiment analysis in sports science: Applications and challenges. *International Journal of Sports Science & Coaching, 16*(3), 251–260. https://doi.org/10.1123/ijsc.2022-0123

[48] Nowak, A., & Dabrowski, M. (2022). Sentiment analysis of political discourse: A comparative study. *Social Research, 89*(2), 45–67. https://doi.org/10.14746/sr.2022.89.2.03

[49] Chen, X., & Wang, Y. (2024). Advances in sentiment analysis for social media data. *Information, 15*(4), 200. https://doi.org/10.3390/info15040200

10.48047/jocaaa.2024.33.08.473

[50] Karami, A. (2021). Natural language processing and sentiment analysis in public health: A systematic review. *Journal of Data and Information Science, 6*(2), 1–25. https://scholarcommons.sc.edu/context/libsci_facpub/article/1290/viewcontent/karamiarticle.pdf

[51] Smith, J., & Brown, K. (2023). Sentiment analysis in healthcare: A deep learning approach. *Journal of Medical Internet Research, 25*(1), e12345. https://doi.org/10.2196/12345 (PMCID: PMC9704505)

[52] Harnessing Topic Modeling and Knowledge Graphs for Enhanced Discovery in Health Interoperability Research Through Retrieval-Augmented Generation. Lee, Adam Micheal(2025). The University of North Carolina at Chapel Hill ProQuest Dissertations & Theses, 2025. 31938573. https://gateway.proquest.com/openurl?res_dat=xri%3Apqm&rft_dat=xri%3Apqdiss%3A31938573&rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Adissertation&url_ver=Z39.88-2004

[53] Tsekouropoulos, G., Vasileiou, A., Hoxha, G., Theocharis, D., Theodoridou, E., & Grigoriadis, T. (2025). Leadership 4.0: Navigating the Challenges of the Digital Transformation in Healthcare and Beyond. *Administrative Sciences*, *15*(6), 194. https://doi.org/10.3390/admsci15060194

[54] Kasperiuniene, J., Briediene, M., Zydziunaite, V. (2020). Automatic Content Analysis of Social Media Short Texts: Scoping Review of Methods and Tools. In: Costa, A., Reis, L., Moreira, A. (eds) Computer Supported Qualitative Research. WCQR 2019. Advances in Intelligent Systems and Computing, vol 1068. Springer, Cham. https://doi.org/10.1007/978-3-030-31787-4_7.

[55] Double-Edged Discourse: The Paradoxical Impact of ESG Terminology on Corporate Governance and Risk Scores Cuartas, Jesus Bill. Drexel University ProQuest Dissertations & Theses, 2024. 31333692. https://www.proquest.com/docview/3077080557

[56] Vishnu S. Pendyala, Karnavee Kamdar, Kapil Mulchandani Automated Research Review Support Using Machine Learning, Large Language Models, and Natural Language Processing. Electronics 2025, 14(2), 256; https://doi.org/10.3390/electronics14020256

[57] Lee, Adam Micheal, Harnessing Topic Modeling and Knowledge Graphs for Enhanced Discovery in Health Interoperability Research Through Retrieval-Augmented Generation. The University of North Carolina at Chapel Hill ProQuest Dissertations & Theses, 2025. 31938573. https://www.proquest.com/docview/3205838403