iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66251

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

The Optimal Number of Topics Detection and their Assessment in Multiple Current Events and Trends Based Datasets Using Topic Models

Gumdelli Manasa[1], Nida Mustafa[2], Muddini Revanth[3], Ch Prabhavathi[4]
[1, 2, 3]*B.Tech, Student, Dept of CSE, Institute of Aeronautical, Engineering Hyderabad , India*
[4]*Assistant Professor, Dept of CSE, Institute of Aeronautical, Engineering Hyderabad , India*

Abstract: ***This project explores the application of sophisticated topic models to determine the ideal amount of subjects and evaluate them across several datasets pertaining to trends and current events. We propose a comprehensive approach that leverages the topic models such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), Probabilistic Latent Dirichlet Allocation (PLDA) and Probabilistic Latent Semantic Indexing (PLSI) to accurately identify and analyse topics. Our approach utilizes various topic modelling techniques to enhance the robustness and precision of topic detection. The results demonstrate the effectiveness of topic models in capturing nuanced trends and themes within complex datasets. A user-friendly frontend, built with Flask and SQLite, enables interactive input and visualization of topic modeling results, offering an accessible platformfor real-time topic analysis and insights. Thismethodology provides a powerful tool for understanding and responding to emerging trends in dynamic datasets.***

Keywords: Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), andLatent Semantic Analysis (LSA), Probabilistic Latent Dirichlet Allocation (PLDA), and Probabilistic Latent Semantic Indexing (PLSI), Topics Modelling, Twitter Dataset

I. INTRODUCTION

Twitter has developed into a vital resource for current information, generating a staggering 12 terabytes of data daily through its 140-character messages [1]. With over 133 million users engaging on the platform, Twitter facilitates discussions across a vast domains, making it a rich resource for textual data. However, the dynamic and noisy nature of this data presents significant challenges for analysis [2]. As researchers and analysts attempt to make sense of this enormous volume of information, effective methods for extracting meaningful insights become crucial.

Topic modelling is a prominent technique used to uncover latent thematic structures within large datasets like those from Twitter. This approach involves identifying clusters of related words that together represent underlying topics within the text [3]. However, choosing the right subjects is one of the most important in topic modeling. This decision directly influences the results' interpretability and use. An optimal number of topics balances granularity with clarity, revealing themes that are both informative and manageable. Conversely, choosing an inappropriate number of topics can lead to either a major piece of information being lost or themes becoming too wide or disjointed [4].

Historically, manual inspection or heuristics have been used to determine number of themes. [5]. These methods often requires more time, potentially leads inconsistent or suboptimal results. As the volume and complexity of data grow, these traditional approaches become increasingly inadequate, necessitating more sophisticated techniques for optimal topic detection.

To overcome these obstacles,this investigates the application of models for detecting and assessing the ideal number within various datasets. Subjectmodels combines strengths of probabilistic models and clustering techniques, offering a more nuanced approach to analyzing Twitter data. By combining content-based data and combined user interaction network structure, these models provide a representation of social connections among individuals as well as timely content, [6]. This integration enhances the ability to identifymeaningful themes while considering the dynamic and interconnected nature of Twitter's information landscape.

The process reuires rigorous assessment and contrast of various arrangements[7]. This approach is essential for applications in advanced and emergingfields, where the quality and accuracy of topic modelling can significantly impact the insights derived from the data. By leveraging topic models, researchers can improve the precision and relevance of topic detection, leading to more actionable and informative results.

©IJRASET: All Rights are Reserved

In summary, as Twitter continues to be a major source of real-time data, effective analysis through advanced topic modelling techniques becomes increasingly important. The development and application of topic models offer a promising solution to the challenges of determining the optimalnumber of topics, ultimately enhancing our ability toextract valuable insights from the vast and complexinformation available on the platform.

II. LITERATURE SURVEY

The analysis of textual data from platforms likeTwitter presents both opportunities and challenges of the enormous amount of the information produced. As Twitter produces approximately 12 terabytes of data daily through its 140-character messages, it has become essential to develop effective methods for extracting meaningful insights from this vast reservoir of information [1]. Topic modelling, a method for locating hidden structures in text data, a powerful tool for navigating the complexities of such large datasets. This literature survey explores various approaches and methodologies related emphasizing the significance of choosing ideal quantity of subjects for effective data analysis.

The concept involves identifying the fundamental ideas or subjects present in a collection of texts. Senthilkumar, Srivani, and Mahalakshmi [3] highlight the use of document topic models to generate word clouds, which visually represent the frequency and significance of words associated with different topics. Word clouds can aid in understanding the thematic structure of a dataset by providing a graphical depiction of key terms. This approach illustrates how topic models can facilitate the visualization of textual data, facilitating consumers' ability to interpret and evaluate outcomes.

However, most important problem is figuring ideal amount of topics. The readability is greatly impactedby the choice of a suitable number of themes. and usefulness of the results. Georgos Balkas, Masih- Amni, and Mariane Clasel [4] tackle this problem byputting forth a topic model created especially for text. It underscores the necessity of choosing a topicnumber that balances the granularity of themes withthe clarity of interpretation. An inappropriate number of topics can lead to either overly broad or fragmented themes, potentially obscuring valuable insights.

Traditional methods for choosing the subjects, such as examination or impromptu heurstics, frequently subjective and time-consuming [5]. They may not always provide consistent or optimal results, prompting the need for more sophisticated techniques. Woer Atevelt, and Welbert [7]emphasize importance for examination of subject modeling. Their research on the examination of a substantial volume of materials gives how systematic evaluation of different topic configurations can enhance the accuracy and relevance of the results. By employing quantitative methods, researchers can objectively assess varioustopic models and identify the most effective subjecs.

Topic models offer a viable remedy for the problemsassociated with topic modelling. Li, Shing, Yaan [10] illustrating how topic approaches can improve the classification of text data by integrating content- based information with additional contextual factors. This makes a nuanced understanding and relevant contextual information.

Rajendar Prasad, Mulana Mohamed [13] explore using vision data highlighting the benefits of visualizing topic models to enhance clustering performance. Their work emphasizes how visual representations of topics can aid in the interpretation and analysis of complex datasets. Visualization tools and techniques play a crucial role in making topic modelling results more accessible and comprehensible, particularly when dealing with large and multifaceted datasets.

The need for advanced methods to ascertain the ideal quantity is further addressed [14]. Their study focuses on the visualization and performance measures for determining the quantity of subjects. They propose a method that combines various methods to increase precision and relevance. This approach highlights the importance of integrating multiple methodologies to address the challenges of topic selection and enhance the overall effectivenessof topic modelling.

In addition to topic models, advancements in topic model visualization have also contributed to improving the interpretation of results. Sergay Karpovech, Smirnav,[15] discuss the use of IPython for topic model visualization. Their work demonstrates how interactive visualization tools can facilitate the exploration and understanding of topicmodels, providing users with a more intuitive means of analyzing complex data. Such tools are invaluable for interpreting the results of topic modelling and making informed decisions based on the insights gained.

In summary, the field of topic modelling has evolved significantly, with various approaches and methodologies being developed to address the challenges of analyzing large and dynamic datasets. It remains a aspect of, influencing the practicality of the findings. Advances in topic models, quantitative analysis, visualization techniques, and interactive tools have all contributed to improval of the relevance of it. As the amount of data keeps increasing, these advancements will play a critical role in enhancing our ability to extract meaningful insights from complex textual data sources like Twitter.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. METHODOLOGY

A. Proposed Work

This project proposes advanced topic modelling systems designed to detect and assess the ideal quantity of subjects related to current events and trends. The system integrates various advanced topic modelling methods, such as Probabilistic Latent Semantic Analysis, Non-negative Matrix Factorization, Latent Dirichlet Allocation, and Latent Semantic Analysis Indexing (PLSI), and Probabilistic Latent Dirichlet Allocation (PLDA). By employing these methods in tandem, the system aims to improve detection's precision. The project features a robust backend for processing and modelling data, and a user-friendly frontend developed with Flask and SQLite, which allows users to input queries, preprocess data, and visualize topic modelling results. This integrated approach ensures a comprehensive analysis of trends and themes, facilitating better insights and understanding of complex datasets.

B. System Architecture

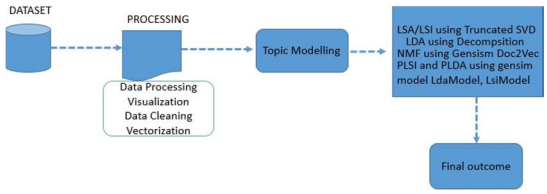


Fig 1. Proposed Architecture

The proposed architecture begins with the ingestion of the dataset, which undergoes a series of preprocessing steps including data processing, visualization, data cleaning, and vectorization. These steps prepare the data for effective topic modelling by converting raw text to structure. Once processed, it is subjected to various topic modelling techniques. LSA and LSI implemented using SVD, while LDA and NMF are applied using Gensim. Additionally, Doc2Vec is utilized to capture document-level semantics, and PLSI along with PLDA are employed through Gensim models. The results from these different models are then analyzed to produce a comprehensive final outcome, highlighting the most relevant topics and patterns within the dataset.

C. Dataset Collection

The dataset collection for this study includes health-related Twitter datasets covering 20 conditions such as Heart Attack, Diabetes, and Breast Cancer, along with TREC2014 topics like Muscle Pain from Statins and Whooping Cough Epidemic, and TREC2015 subjects including Paradox Side Effects and Lyme Disease Study. Additionally, datasets from TREC2018 and COVID-19 were incorporated. Key words from these datasets were utilized for experimentation, with implementation in Python and MATLAB environments.



	SNO	Tweet	Topic
0	1	TRENDING: New Yorkers encounter empty supermar...	Covid
1	2	#Panic buying hits #NewYork City as anxious sh...	Covid
2	3	#toiletpaper #dunnypaper #coronavirus #coronav...	Covid
3	4	Do you remember the last time you paid $2.99 a...	Covid
4	5	@DrTedros "We can't stop #COVID19 without pro...	Covid

Fig 2 Dataset

D. Processing

1) Data Processing

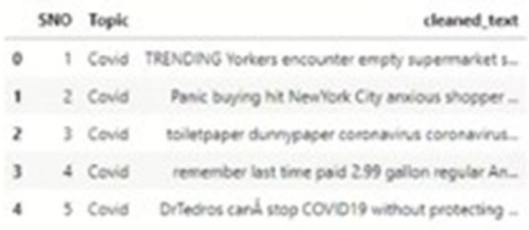
Data processing involves preparing raw datasets for analysis by structuring and organizing the data. This step includes tasks such as parsing, sorting, and transforming data into a format that facilitates efficient analysis. The goal is to ensure that the data is accurately represented and ready for subsequent stages, such as visualization and modeling. Effective data processing is crucial for extracting meaningful insights from complex datasets.

2) Visualization

Visualization using Seaborn and Matplotlib involves creating graphical representations of data to uncover patterns and insights. Seaborn, built on Matplotlib, provides advanced visualization features and aesthetically pleasing charts, while Matplotlib offers a broader range of customizable plots. Together, these libraries enable the creation of detailed plots such as heatmaps, bar charts, and scatter plots, which help in understanding data distributions and relationships.

3) Data Cleaning

Finding and fixing errors or inconsistencies in the data, managing , eliminating duplication, standardizing formats are all part of this process. By ensuring that the data is accurate and consistent, data cleaning improves the quality of the dataset and enhances the reliability of the analysis. Proper data cleaning is essential for producing valid andactionable insights from the data.



	SNO	Topic	cleaned_text
0	1	Covid	TRENDING Yorkers encounter empty supermarket s...
1	2	Covid	Panic buying hit NewYork City anxious shopper ...
2	3	Covid	toiletpaper dunnypaper coronavirus coronavirus...
3	4	Covid	remember last time paid 2.99 gallon regular An...
4	5	Covid	DrTedros canÂ stop COVID19 without protecting ...

Fig 3 Data Cleaning

4) Vectorization of the Data

Vectorization of the data transforms textual information into numerical vectors suitable for machine learning models. Techniques such as Bag of Words, TF-IDF, and word embeddings are used to convert text into vectors that represent the semantic meaning of words. This process allows algorithms to process and analyze textual data efficiently, enabling effective modelling and pattern recognition in various applications.

E. Topic Modelling

1) LSA/LSI using TruncatedSVD:

LSA and LSI leverage Truncated Singular Value Decomposition to uncover latent topics in textual data. This process involves transforming the data to space with fewer dimensions, each of which represents a hidden topic. Documents and topics numbers is specified to define the granularity of the analysis. TruncatedSVD reduces dimensionality by retaining only the most significant singular values, thus capturing the core semantic structure of the data. After transformation, the model determines which words are most crucial to each issue., highlighting key terms that best represent the underlying themes. This approach enables efficient topic modelling by focusing on the most relevant features, facilitating the extraction of meaningful insights and thematic patterns from large text corpora.

2) LDA using Decomposition

Latent Dirichlet Allocation (LDA) with decomposition techniques models the topics within textual data by presuming that all documents are collections of themes, and that all topics are collections of words. Process begins in formatting the data appropriately, indicating the quantity of documents and subjects that will be the focus of theanalysis. It decomposes the document-term matrix into topic distributions, revealing how documents are composed of various topics. For instance, the composition of Document 0 is determined by the relative weights of each subject in it . Also, LDA identifies the most important words for each topic, ranking terms based on their relevance and contribution to the topic's characterization. This approach provides a detailed understanding of document content and the thematic structure of the dataset, facilitating effective interpretation andanalysis of large text corpora.

3) NMF using Gensim Doc2Vec

Non-negative Matrix Factorization (NMF) combined with Gensim's Doc2Vec model processes textual data by first converting documents into vector representations using Doc2Vec. This involves vectorizing the data, storing abstracts in a 'data' repository, and transforming the information into compatible list structure. One function is employed to display the topics derived from the model.

The NMF model is built from these vectors and saved to a file for future use. Once the model isloaded, it allows for the extraction of topics based on a specific document number, providing insights into the thematic content of individual documents. This approach enables effective topic modeling by leveraging document embeddings to capture semantic relationships, facilitating a better comprehension of the dataset's fundamental subjects.

4) PLSI and PLDA using Gensim Models LdaModel, LsiModel

PLSI and PLDA are implemented using Gensim's LdaModel and LsiModel for topic modeling. The process begins with data processing and vectorization to convert textual information into numerical format. Once the data is prepared, the models are loaded and applied to uncover topics. The LdaModel facilitates PLDA, while the LsiModel is used for PLSI. Each model extracts topics from the data, and the results are examined by printing the topics along with their most significant words. This allows for a comprehensive understanding of the thematic structure, as PLSI and PLDA reveal different aspects of the data's latent topics. By leveraging Gensim's capabilities, this approach effectively captures and interprets the key themes within the dataset, providing valuableinsights into the underlying content.

IV. EXPERIMENTAL RESULTS

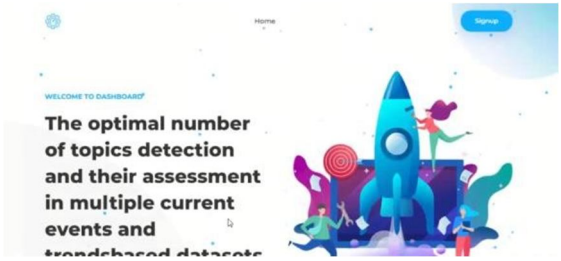


Fig 4 Home Page

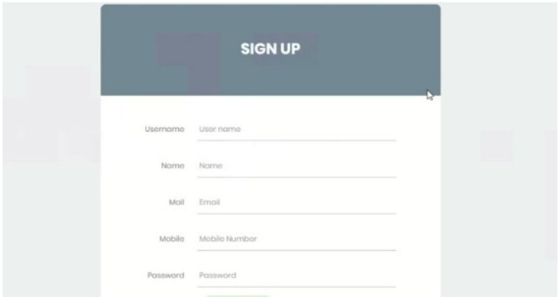


Fig 5 Signup Page



Fig 6 Signin Page

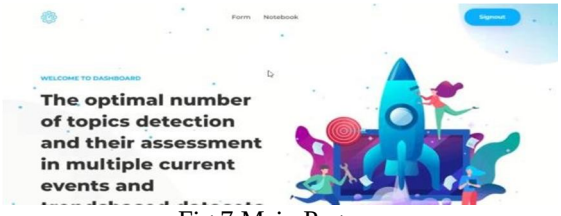


Fig 7 Main Page



Fig 8 Input Page



Fig 9 Output Screen



Fig 10 Input Sample Excel



Fig 11 Input Screen

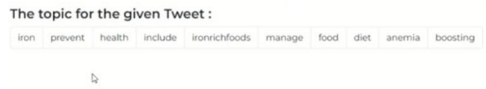


Fig 12 Predicted Results

V. CONCLUSION

The project successfully demonstrates the efficacy of topic modeling techniques in detecting and analyzing topics within datasets related to current events and trends. By developing Probabilistic Latent Semantic Indexing (PLSI), Probabilistic Latent Dirichlet Allocation (PLDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), system offers a robust framework for uncovering nuanced patterns and themes across diverse datasets. The comparative analysis of these methods reveals their complementary strengths, enhancing the overall accuracy and depth of topic detection. The implementation of a user-friendly frontend using Flask and SQLite provides an accessible interface for users to interact with the system, facilitating real-time data input, preprocessing, and visualization. This models bridges the gap between complex topic modeling processes and practical, actionable insights.

Overall, the project highlights the potential of topic modeling approaches in advancing our understanding of dynamic trends and emerging themes. The system not only improves topicdetection but also offers a practical tool forstakeholders to engage with and analyze contemporary data effectively.

VI. FUTURE SCOPE

Future scope includes enhancing the system with advanced deep learning models for improved topic detection and classification accuracy. Integrating real-time data streams and multilingual support will broaden its applicability. Additionally, expanding the system to handle larger, more diverse datasets will further refine its analysis of complex trends. Collaborations with domain experts could enable customized insights, making the tool invaluable for researchers, analysts, and decision-makers across various industries.

REFERENCES

[1] ShaohuaLi & Tat-SengChua (2017). Document Visualization using Topic Clouds. Informationretrieval, DOI: arXiv:1702.01520v1.
[2] Xu G., Meng Y., Chen Z., Qiu X., Wang C., & Yao H (2019). Research on Topic Detection and Tracking Online News Texts. IEEE Access, 7,58407–58418. DOI:10.1109/access.2019.2914097.
[3] Senthilkumar S, Srivani M & Mahalakshmi G.S. (2017, June). Generation of the word clouds using Document Topic Models. Conference paper on Second International Conference in Recent Trends and Challenges in Computational Models. DOI:10.1109/ICRTCCM.2017.60.
[4] Georgios Balikas, Massih-Reza Amini, & Marianne Clausel (2016, July). On a Topic Model for Sentences. Conference paper in proceedings of 39th International ACM SIGIR Conference onResearch and Development in InformationRetrieval. https://doi.org/10.1145/2911451.2914714.
[5] Pattanodom, M., I am-On, N., & Boongoen, T (2016, January). Clustering data with presence of missing values by ensemble Approach. Conference paper on 2016 second Asian Conference in DefenseTechnology (ACDT), Thailand.DOI:10.1109/acdt.2016.7437660.
[6] Rajendra Prasad K, Moulana Mohammed &Noorullah R M (2019). Hybrid Topic Cluster Models for Social Healthcare Data. InternationalJournal of Advanced Computer Science and Applications (IJACSA), 10(11), 491-506. DOI: 10.14569/IJACSA.2019.0101168.
[7] Carina Jacobi, Wouter van Atteveldt & Kasper Welbers (2015). Quantitative analysis for large amounts of journalistic texts using topic modelling. Digital Journalism. DOI10.1080/21670811.2015.1093271.
[8] Damir Korenci C, Strahil Ristov, & Jan Sneijder (2018). Document-based measure of Topic Coherence for News Media Text.Expert Systems with Applications, 114, 357-373. DOI:10.1016/j.eswa.2018.07.063.
[9] Carson Sievert & Kenneth E. Shirley (2014, June). LDAvis: A method of visualizing and interpreting topics. Conference paper in Proceedings of the Workshop on Interactive Language Learning,Visualization, and Interfaces, Baltimore, Maryland,USA. DOI:10.3115/v1/W14-3110.
[10] Li Z., Shang W, & Yan M. (2016, June). Newstext classification model based on the topic model, IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), DOI:10.1109/icis.2016.7550929.
[11] Alessia Amelio & Clara Pizzuti (2016, February). Is Normalize Mutual Information Fair Measure to Compare Community Detection Methods? Conference paper in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France. DOI:10.1145/2808797.2809344.
[12] Dataset keyword phrases 2018 PrecisionMedicine Track TREC2018. Track web pagehttp://www.trec-cds.org/.
[13] Rajendra Prasad K, Moulana Mohammed & Noorullah R M (2019). Visual Topic Models for Healthcare Data Clustering.Evolutionary Intelligence, 1, 1-17. https://doi.org/10.1007/s12065-019-00300-y.
[14] Noorullah, R.M. and Mohammed, Moulana. 'Visualization and Performance Measure to Determine Number of Topics in Twitter Data Clustering Using Topic Modeling'. 1 Jan. 2021: 803– 817. DOI: 10.3233/JIFS-202707.
[15] Sergey Karpovich, Alexander Smirnov, Nikolay Teslya & Andrei Grigorev (2017, October). Topic Model Visualization with (14) IPython. Conference paper in proceeding of the 20th conference of Fruct Association. DOI: 10.23919/FRUCT.2017.8071303.
[16] Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). Finding Natural Number of Topics in Latent Dirichlet Allocation: Some Observation. Pacific-Asia Conference inKnowledge Discovery and Data Mining. DOI: 10.1007/978-3-642-13657-3_43.
[17] Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science. The Annals of Applied Statistics, 1(1), 17-35. DOI: 10.1214/07- AOAS114.
[18] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interprets Topic Models. Advances Neural Information Processing Systems, 22.
[19] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal Machine Learning Research, 3, 993-1022.
[20] Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics. Proceeding of the National Academy of Sciences, 101(Suppl 1), 5228-5235. DOI: 10.1073/pnas.0307752101.
[21] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. Proceedings of the Conference on Empirical Methods in NaturalLanguage Processing. DOI:10.5555/2145432.2145462.
[22] Blei, D. M., & McAuliffe, J. D. (2007). Supervised Topic Modeling. Advances Neural Information Processing Systems, 20.
[23] Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improval of Topic Coherence withRegularized Topic Models. Advance in Neural Information Processing Systems, 24.
[24] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures.Proceeding of Eighth ACM International Conference on Web Search and Data Mining. DOI: 10.1145/2684822.2685324.
[25] Stevens, K., Kegelmeyer, P., Andrzejewski, D.,& Buttler, D. (2012). Exploring of Topic Coherenceon Many Models and Many Topics. Proceedings of the 2012 Joint Conference in Empirical Methods in Natural Language Processing and ComputationalNatural Language Learning. DOI:10.5555/2390948.2391076.
[26] Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. Foundation and Trends in Information Retrieval, 11(2-3), 143-296. DOI:

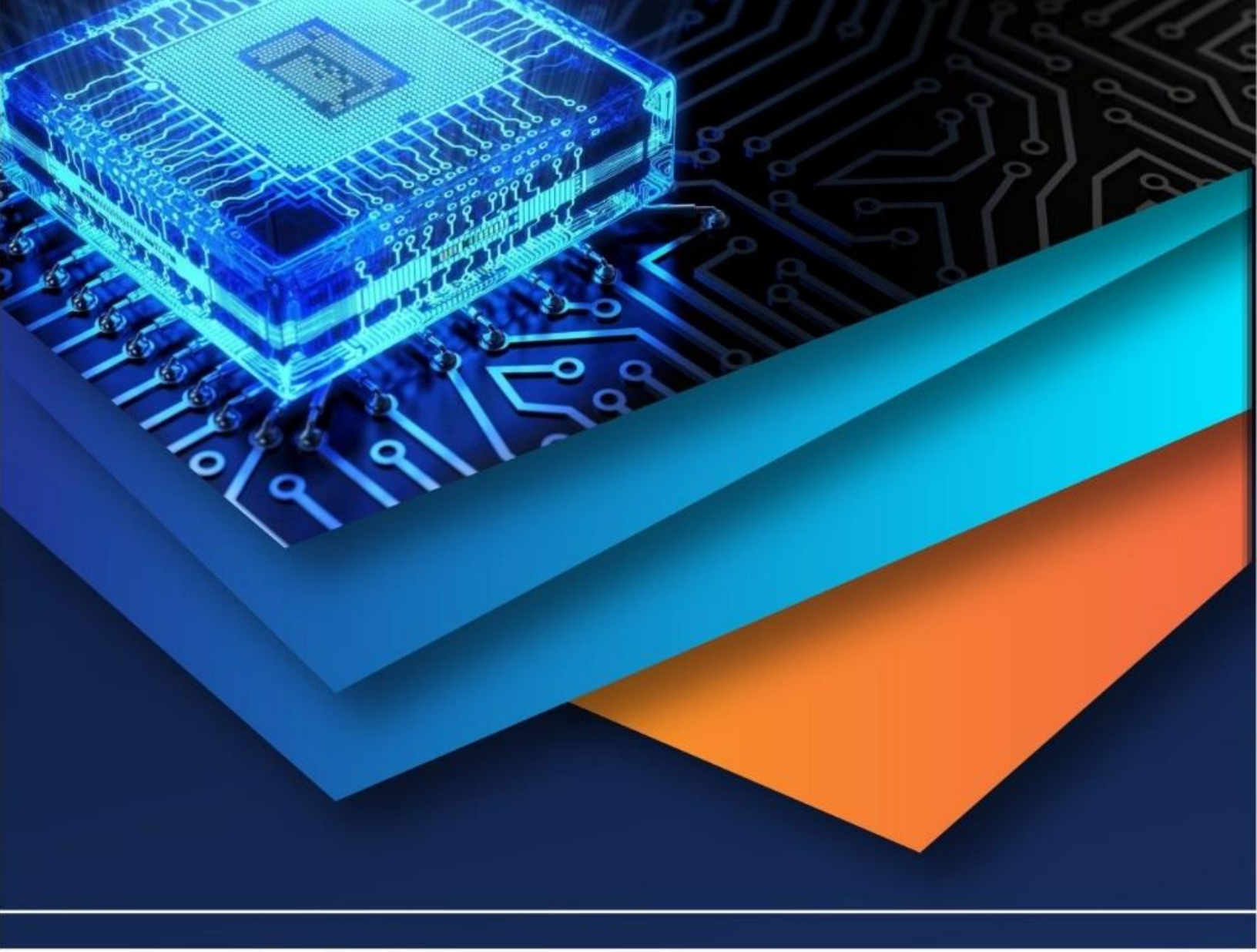
10.1561/1500000030.

[27] Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), 1064-1082. DOI: 10.1111/ajps.12103.

[28] Silge, J., & Robinson, D. (2017). Text Mining with R: A Tidy Approach. O'Reilly Media.

[29] Alghamdi, R., & Alfalqi, K. (2015). Survey of Topic Modelling in Text Mining. International Journal in Advanced Computer Science and Applications (IJACSA), 6(1).

[30] Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation Methods for Topic Models. Proceeding of 26th Annual International Conference on Machine Learning. DOI:10.1145/1553374.1553515.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)