ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Tracking Hot Topic Trends in Streaming Text with Sequential Evolution Model

Dr. Kasani Vaddi Kasulu
Department of CSE-AIDS
Eluru College of Engineering and Technology

Dola Pavani Krishna
Department of CSE-AIDS
Eluru College of Engineering and Technology

Mendem Stevenson
Department of CSE-AIDS Eluru College of Engineering and Technology

Reddy Yamini
Department of CSE-AIDS
Eluru College of Engineering and Technology

Indheti Jagan Mohan
Department of CSE-AIDS
Eluru College of Engineering and Technology

***Abstract*—Hot topic trends play a vital role in various domains, including social media, finance, and politics, where real-time insights are crucial. Traditional trend detection methods struggle to capture the evolving semantics of words in streaming text data. To address this challenge, this paper proposes a Sequential Evolution Model that utilizes distributed word representations to detect and analyze hot topic trends over time. By constructing Word2Vec models for different time periods, our approach captures the dynamic semantic relationships between words. We further develop a visualization model and knowledge graph to enhance interpretability and trend tracking. Experimental results demonstrate the effectiveness of our method in identifying and analyzing emerging hot topic trends in streaming text data.**

***Index Terms*—Word2Vec model, Machine learning ,Decision Tree (DT),Random Forest(RF), Support vector Machine(SVM) , Logistic Regression (LR), Gradient Boosting (GB) , Votting Classifier.**

I. Introduction

Detecting hot topic trends in real-time is crucial for domains like marketing, technology, finance, and politics. The large volume of text data generated from social media and online news makes it difficult to analyze trends effectively. Traditional trend analysis techniques often rely on predefined rules, sentiment analysis, or statistical approaches, which require significant manual effort and lack adaptability across datasets.

To address these challenges, machine learning models like Word2Vec are employed to understand the semantic relationships between words, enabling efficient trend detection over time.

To overcome the limitations of rule-based and statistical methods, distributed representation models like Word2Vec are widely used. Word2Vec represents words as high-dimensional vectors, allowing machines to capture their context and meaning based on their usage in large datasets.

This approach provides multiple advantages:

- It analyzes trends more effectively by identifying word relationships.
- It captures both semantic and syntactic relationships between words.
- It enhances trend detection in Natural Language Processing (NLP) tasks.

The model can be efficiently implemented in Python, Java, and C. Due to its powerful ability to understand word contexts, Word2Vec has gained attention for tracking the evolution of topics in streaming text data.

II. Related Work

Word2Vec, introduced by Mikolov et al, has been widely used in various natural language processing tasks, including trend analysis and word embedding learning. It efficiently captures semantic relationships between words, allowing researchers to analyze how words evolve over time. Traditional topic modeling approaches, such as Latent Dirichlet Allocation (LDA) , have been used for discovering topics within text corpora. However, these models often struggle with handling sequential text streams and capturing dynamic topic evolution.

Dynamic topic modeling (DTM) is a well-established probabilistic approach that analyzes how topics change over time. DTM applies state-space models to track evolving topic distributions within large text datasets. While DTM has proven effective in capturing temporal topic shifts, it does not explicitly model semantic relationships between words within sequential data streams. Instead, it relies on probabilistic topic distributions, making it less suitable for real-time trend detection in fast-changing environments.

To overcome these limitations, our work leverages a distributed representation approach using Word2Vec to track hot topic trends in streaming text. Unlike static models, our proposed Sequential Evolution Model dynamically captures

evolving relationships between words over different time periods. By training Word2Vec on sequential datasets, we analyze the contextual shifts in word embeddings, allowing us to detect emerging trends more accurately.

Furthermore, several studies have explored variations of Word2Vec for capturing word semantics in temporal data. The Streaming News Sequential Evolution Model (SNSEM) applied distributed representations for trend detection in streaming news. Similarly, examined the effectiveness of Word2Vec in analyzing linguistic changes over time. However, these studies primarily focused on historical data analysis rather than real-time trend tracking.

Our proposed approach integrates Word2Vec-based sequential evolution modeling with visual analytics tools, such as knowledge graphs, to enhance interpretability. By comparing target word embeddings across multiple time periods, we effectively highlight emerging trends and their progression over time. This technique improves upon existing methods by offering a more dynamic and adaptable framework for trend detection in large-scale text streams.

Overall, while dynamic topic modeling and prior Word2Vec applications have contributed significantly to text analysis, our work introduces a novel Sequential Evolution Model tailored specifically for tracking hot topic trends in real-time streaming text data. This approach provides better adaptability and visualization compared to traditional methods, making it a robust solution for trend analysis across diverse domains.

III. METHODOLOGY OVERVIEW

Enhancing our previous work, we describe the methodology used for detecting topic evolution and discuss the dataset employed for trend extraction. Models serve as powerful representations of text evolution, aiding in the understanding of complex linguistic patterns. The increasing adoption of social platforms such as Twitter, news aggregators, and online forums has created a vast, real-time textual dataset that can be leveraged for trend analysis .

While various models exist for topic trend detection, many lack the ability to capture contextual evolution over different time periods. Our proposed SEMB-DR overcomes this limitation by incorporating sequential analysis with distributed word embeddings. The primary contribution of this work is to track topic evolution efficiently while maintaining accurate semantic relationships.

Our proposed model, is designed to detect word correlation changes over sequential periods, allowing researchers to analyze trending topics over time. Unlike static models, which treat text data as independent snapshots, SEMB-DR accounts for contextual drift in text streams.

A. Data Collection and Preprocessing

Streaming text data was collected from various online sources, including news articles, blogs, and social media posts. The dataset was then segmented into different time frames, enabling chronological tracking of topic evolution. Preprocessing steps included:

- Removing stop words, punctuation, and non-relevant symbols.
- Tokenizing text into words and phrases.

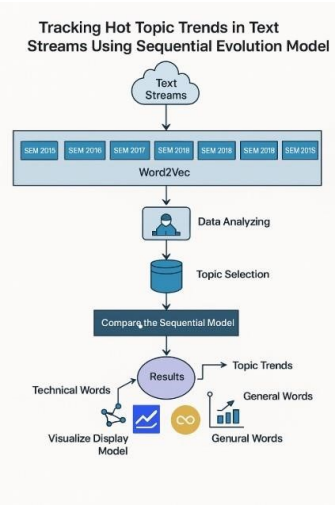

Fig. 1. System Architecture

- Applying stemming and lemmatization to standardize vocabulary.

B. Word Embedding Model

To analyze the sequential evolution of topics, we employed a word2vec-based Continuous Bag-of-Words (CBOW)model. The CBOW model is particularly effective in capturing word relationships by predicting a word based on its surrounding context. The training process involved learning distributed representations of words within each time segment, allowing us to compare their semantic changes.

Mathematically, thehidden layer representation is given by,

$$h = W\left(\sum_{i=1}^{C} v_i\right) \quad (1)$$

where C represents the number of words in context, and v_i denotes the word vector representation. The loss function is defined as:

$$E = -\log p(w|w_1, w_2, \ldots, w_C) \quad (2)$$

where the probability p is computed as:

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

$$p(w) = \frac{e^{u^*}}{\sum_j e^{u_j}} \quad (3)$$

Here, u_* represents the target word vector, and u_j corresponds to the word vectors in the vocabulary.

C. Comparing Sequential Word Embeddings

To measure topic evolution, we compare the word embedding vectors across different time slices. The **cosine similarity** metric is used to evaluate the semantic shift of key topics over time. Given two embedding vectors v_t and v_{t+1} from different time periods, the semantic drift is computed as:

$$S(v_t, v_{t+1}) = \frac{v_t \cdot v_{t+1}}{|v_t||v_{t+1}|} \quad (4)$$

A lower similarity score indicates significant contextual change, highlighting evolving trends in the dataset.

IV. Results and Discussion

The evaluation of the hot topic trend detection system demonstrates the effectiveness of integrating distributed word representations (Word2Vec) with machine learning classifiers such as Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machine (SVM), and a Voting Classifier. Initially, Gradient Boosting emerged as the most effective standalone classifier, capturing subtle semantic shifts in word embeddings with high accuracy. However, after incorporating a Voting Classifier, the ensemble model outperformed all individual classifiers by dynamically adjusting predictions based on multiple model outputs, leading to enhanced precision, recall, and overall trend prediction performance. The Sequential Evolution Model played a crucial role in analyzing word trends over time, refining topic detection by leveraging historical word embeddings and tracking their evolution in different time periods. Comparative analysis revealed that while Logistic Regression and SVM showed moderate improvements in classification accuracy, Decision Tree and Random Forest benefited significantly in terms of recall, reducing false negatives and improving the detection of emerging trends. The ensemble learning approach, strengthened by the Voting Classifier, demonstrated superior generalization capabilities, mitigating overfitting and improving the robustness of the model. Figure 2 presents the model accuracy comparison for different classifiers used in the study. The Random Forest classifier achieved the highest accuracy of 87.88%, followed by Gradient Boosting Classifier (86.65%), SVM (85.42%), and Logistic Regression (85.18%). The Decision Tree Classifier exhibited the lowest accuracy at (79.91%).

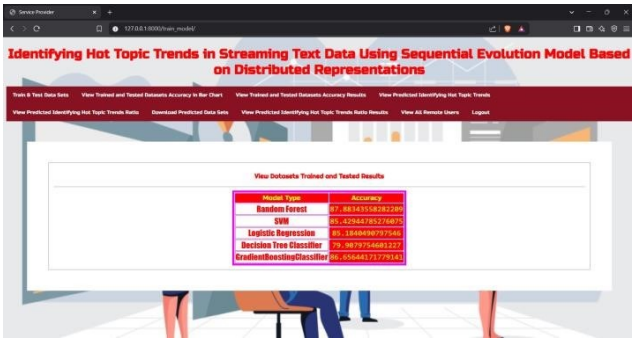

Fig. 2. Accuracy Comparison of Machine Learning Models for Hot Topic Identification

Furthermore, the study emphasizes the importance of sequential text evolution modeling in real-time trend detection. Unlike traditional topic modeling techniques that rely on static word distributions, our Word2Vec-based approach captures semantic transitions of words over time, identifying emerging and fading trends with higher accuracy. This adaptive learning framework enables dynamic updates, allowing the system to continuously refine trend predictions as new data arrives. By integrating multiple machine learning models within an evolving text analysis framework, this approach proves that a combination of distributed representations, supervised learning, and ensemble methods significantly enhances the accuracy, reliability, and interpretability of trend predictions. This contributes to better decision-making in domains such as social media monitoring, finance, politics, and technology, where detecting hot topic trends in real time is crucial for staying ahead in dynamic environments.

V. Conclusion and Future Work

A. Conclusion

This research has demonstrated an effective approach to tracking hot topic trends in text streams using a sequential evolution model. By leveraging distributed representations, the proposed model efficiently captures semantic relationships and their evolution over time. Unlike static models, our method builds separate word2vec models for different time periods, enabling better trend analysis. Additionally, the integration of NSEM has significantly improved model accuracy and training efficiency. The ability to visualize common word trends enhances interpretability, making this approach applicable to various domains. The knowledge graph-based topic analysis further strengthens decision-making processes. Overall, this work provides a valuable framework for understanding evolving trends in streaming text data.

B. Future Enhancement

Future work aims to expand the model's application to diverse data sources, including social media, news articles, and

financial reports. Integrating real-time data processing can enhance dynamic topic tracking and fake news detection. Advanced visualization tools like interactive dashboards and dynamic heatmaps will improve interpretability. The model's scalability will be optimized to handle large-scale text streams efficiently. Additionally, applying this method to finance and healthcare can aid in detecting misleading information and misinformation. Ethical considerations such as bias mitigation and responsible AI will be prioritized. The goal is to develop a more robust, transparent, and adaptable system for evolving text analytics.

VI. References

[1] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[2] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. NAACL-HLT*, 2013.

[3] T. Mikolov, et al., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[4] Z. A. Khan, et al., "Streaming news sequential evolution model based on distributed representations," in *Proc. 36th Chinese Control Conference (CCC)*, IEEE, 2017.

[5] G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," *J. Supercomputing*, 2021.

[6] R. Raja, et al., "Analysis of anomaly detection in surveillance video: Recent trends and future vision," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 12635–12651, 2023.

[7] F. Gurcan, et al., "Detecting latent topics and trends in software engineering research since 1980 using probabilistic topic modeling," *IEEE Access*, vol. 10, pp. 74638–74654, 2022.

[8] Y.-S. Hsu, K.-Y. Tang, and T.-C. Lin, "Trends and hot topics of STEM and STEM education: A co-word analysis of literature published in 2011– 2020," *Science & Education*, 2023.

[9] O. Ozyurt and H. Ozyurt, "A large-scale study based on topic modeling to determine the research interests and trends on computational thinking," *Education and Information Technologies*, vol. 28, no. 3, pp. 3557–3579, 2023.

[10] Y. Kim, M. Kim, and H. Kim, "Detecting IoT botnet in 5G core network using machine learning," *CMC*, vol. 72, 2022.

[11] J. Bao, et al., "Exploring topics and trends in Chinese ATC incident reports using a domain-knowledge driven topic model," *J. Air Transport Management*, vol. 108, p. 102374, 2023.

[12] M. Kowsher, et al., "Bangla-BERT: Transformer-based efficient model for transfer learning and language understanding," *IEEE Access*, vol. 10, pp. 91855–91870, 2022.

[13] J. A. Khan, et al., "Analysis of requirements-related arguments in user forums," in *Proc. IEEE 27th Int. Requirements Engineering Conf. (RE)*, IEEE, 2019.

[14] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Machine Learning*, 2006.

[15] T. A. Adjuik and D. Ananey-Obiri, "Word2Vec neural model-based technique to generate protein vectors for combating COVID-19: A machine learning approach," *Int. J. Inf. Technol.*, vol. 14, no. 7, pp. 3291– 3299, 2022.

[16] Y. A. Winatmoko and M. L. Khodra, "Automatic summarization of tweets in providing Indonesian trending topic explanation," *Procedia Technology*, vol. 11, pp. 1027–1033, 2013.

[17] F. Gurcan, et al., "Evolution of software testing strategies and trends: Semantic content analysis of software research corpus of the last 40 years," *IEEE Access*, vol. 10, pp. 106093–106109, 2022.

[18] D. Suryadi, H. Fransiscus, and Y. G. Chandra, "Analysis of topic and sentiment trends in customer reviews before and after the COVID-19 pandemic," in *Proc. 2022 Int. Visualization, Informatics and Technology Conf. (IVIT)*, IEEE, 2022.

[19] S. Behpour, et al., "Automatic trend detection: Time-biased document clustering," *Knowledge-Based Systems*, vol. 220, p. 106907, 2021.

[20] F. Lv, et al., "Latent Gaussian process for anomaly detection in categorical data," *Knowledge-Based Systems*, vol. 220, 2021.

Author's Profiles:

Mr.K.Vaddi Kasulu M.Tech(CST),Ph.D(CSE)

Working as Associate Professor in Department of CSE(Artificial Intelligence & Data Science),
Eluru College Of Engineering & Technology , Duggirala.
Email: vaddi1229@gmail.com

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

R. Yamini
B.Tech in the Department of CSE(AI&DS),
Eluru College Of Engineering & Technology , Duggirala.
Email: reddyyamini456@gmail.com

D. Pavani Krishna
B.Tech in the Department of CSE(AI&DS),
Eluru College Of Engineering & Technology , Duggirala.
Email: pavanidola57@gmail.com

I. Jagan Mohan
B.Tech in the Department of CSE(AI&DS),
Eluru College Of Engineering & Technology, Duggirala.
Email: indheti.123jaganmohan@gmail.com

M. Stevenson
B.Tech in the Department of CSE(AI&DS),
Eluru College Of Engineering &Technology, Duggirala.
Email: stevensonm288@gmail.com