International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

IDENTIFYING HOT TOPIC TRENDS IN STREAMING TEXT DATA USING SEQUENTIAL EVOLUTION MODEL BASED ON DISTRIBUTED REPRESENTATIONS

[1]ATHINARAPU KIRAN SAGAR,[2]MITTAPELLY SADWIK CHANDRA,[3]AARE RAMYA KRISHNA,[4]MALOTH KRISHNA,[5]DR.K.ARUN KUMAR

[1,2,3,4]Students, Department of computer Science And Engineering,Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100
[5]Professor, Department of computer Science And Engineering,Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100

ABSTRACT

In today's social media-driven world, the rapid dissemination of trending topics significantly influences public discourse and behavior. However, the task of identifying and analyzing these trends from large-scale, real-time text streams presents substantial challenges, especially in capturing the semantic shifts over time. This study proposes an innovative approach that utilizes distributed representations to detect and track emerging topics. A sequential evolution model is introduced for a streaming news platform, allowing for the identification of trending topics as they evolve. To facilitate understanding and analysis, the study incorporates both a visual representation model and a knowledge graph. The approach involves gathering streaming news data, segmenting it chronologically, and applying word embedding models to different time intervals. By studying the semantic evolution of key terms, the model effectively uncovers trends and their transformations over time. Experimental results confirm the model's ability to accurately identify hot topics and provide valuable visual insights, surpassing traditional methods in effectiveness.

Keywords: Social Media Trends, Trending Topic Detection, Streaming Text Data, Semantic Evolution, Topic Analysis, Distributed Representations.

INTRODUCTION

Real-time trend analysis has become increasingly important in a wide range of fields, including business, technology, finance, and politics. Understanding and tracking trends helps organizations make informed strategic decisions and adapt to shifting public interests. However, traditional methods of trend detection often struggle with the challenges posed by large datasets and the subtle linguistic changes that occur over time. As a result, there is a growing need for more effective techniques to analyze and interpret streaming text data. Recent advancements in machine learning and natural language processing (NLP) have greatly enhanced our ability to detect semantic relationships and emerging patterns in text. Techniques such as Word2Vec, which use word embeddings to capture contextual meaning, have significantly improved trend detection. These methods enable the identification of latent connections between words and phrases that might otherwise go unnoticed. However, while these developments are promising, they still face limitations in accurately tracking trends over time. One of the key challenges in current trend detection methods is the static nature of many word embedding models. These models typically

rely on fixed representations of words that do not account for changes in meaning or usage over time. As a result, many models struggle to capture the dynamic evolution of language, which is essential for tracking trends as they emerge and transform. Additionally, many existing models lack robust visualization capabilities, which makes it difficult to interpret the results and gain actionable insights from the data. To address these challenges, this study proposes a novel approach based on a sequential evolution model for analyzing streaming text data. The model divides the data into chronological segments, allowing for the detection of emerging topics and their evolution over time. By applying word embedding techniques to these segmented datasets, the model is able to capture the shifts in word relationships and semantic meaning that occur as trends develop. This approach ensures a more accurate and nuanced understanding of the changing nature of trends. In addition to the sequential evolution model, this research incorporates a visual representation model and a knowledge graph to enhance data interpretation. These tools provide a clearer, more intuitive way to explore and analyze the relationships between words and trends. By integrating these visualization tools, the study aims to improve both the accuracy of trend detection and the clarity of the insights generated. Ultimately, this research seeks to contribute to the advancement of NLP techniques and real-time trend analysis by providing more powerful tools for understanding and tracking emerging trends.

LITERATURE SURVEY

T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," Advances in Neural Information Processing Systems, vol. 26, 2013.
In this seminal paper, Mikolov and colleagues introduce the concept of distributed word representations, commonly known as word embeddings, and their ability to capture syntactic and semantic word relationships. The authors propose a model called Word2Vec, which learns vector representations of words from large corpora of text by exploiting their contextual relationships. This work addresses the challenge of understanding word meaning based on context rather than predefined dictionary definitions, highlighting the compositionality property of word embeddings. By demonstrating how words and phrases can be represented as vectors in a continuous space, the study lays the groundwork for various applications in natural language processing, including trend detection, sentiment analysis, and machine translation. This paper has had a profound impact on the development of NLP models and has inspired a wide range of follow-up research on word embeddings and their applications in machine learning.

T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in Proc. 2013 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2013.
Mikolov et al. in this paper focus on uncovering linguistic regularities within continuous space word representations, further expanding on the Word2Vec framework. The authors present experimental evidence showing that word embeddings can capture semantic relationships between words, such as analogy-based regularities (e.g., “king” is to “queen” as “man” is to “woman”). By using

word vectors trained on large datasets, they demonstrate that linguistic relationships can be derived from these embeddings, significantly enhancing the ability to perform tasks such as word analogy and classification. This research helped establish word embeddings as a crucial tool in NLP by showing that machine learning models could learn rich, interpretable features of language directly from text data. The insights provided by this paper have influenced numerous applications, from syntactic parsing to topic modeling and trend detection in dynamic texts.

T. Mikolov, et al., "Efficient estimation of word representations in vector space," arXiv preprinarXiv:1301.3781,2013.
In this paper, Mikolov and his collaborators focus on improving the efficiency of learning word representations in vector space, introducing a new approach for training word embeddings. They propose two techniques: the Continuous Bag of Words (CBOW) and the Skip-Gram model, both designed to predict a word based on its context or vice versa. These methods are computationally efficient, allowing them to scale to massive datasets and generate word vectors that capture semantic meanings and syntactic relationships. The Skip-Gram model, in particular, demonstrated better performance when applied to large corpora. This work is crucial in the context of trend detection, as it allows for the creation of high-quality word embeddings that can be applied to dynamic and real-time data. The methodology presented here continues to form the backbone of many modern NLP models used in diverse applications such as sentiment analysis, topic modeling, and trend analysis.

Z. A. Khan, et al., "Streaming news sequential evolution model based on distributed representations," in Proc. 36th Chinese Control Conference (CCC), IEEE, 2017.
Khan et al. introduce a sequential evolution model designed to analyze streaming news data by leveraging distributed representations of words. Their model aims to capture the evolution of trends and topics in real-time, addressing the challenge of detecting emerging themes in a constantly updating news stream. By applying word embeddings in a sequential manner, the model is able to track changes in semantic relationships over time, which is essential for understanding how topics evolve in dynamic environments like news platforms. The incorporation of time-segmented data helps improve trend detection accuracy by highlighting temporal shifts in language usage. The research showcases the potential of combining word embeddings with sequential models to enhance real-time data analysis and is particularly valuable for applications in social media trend detection, news analysis, and real-time topic modeling.

G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," J. Supercomput., 2021.
Di Gennaro et al. delve into the practical considerations involved in learning Word2Vec embeddings and optimizing their performance for various NLP tasks. They discuss the trade-offs between computational efficiency and the quality of learned word representations, emphasizing the importance of hyperparameter tuning, model selection, and corpus characteristics in determining the effectiveness of word embeddings. The authors provide an in-depth analysis of the challenges encountered when training Word2Vec, such as dealing with sparse data and balancing model complexity with runtime efficiency. The

paper highlights the significant role of Word2Vec in modern NLP applications, from sentiment analysis to trend detection, and provides insights on how to improve its performance in real-world applications. This work is instrumental for researchers and practitioners looking to optimize word embedding models for specific tasks like dynamic trend detection and semantic analysis in large-scale datasets.

R. Raja, et al., "Analysis of anomaly detection in surveillance video: recent trends and future vision," Multimedia Tools Appl., vol. 82, no. 8, pp. 12635-12651, 2023.
Raja et al. review recent advancements in anomaly detection within the context of surveillance video analysis, focusing on the methods and models used to identify unusual events or behaviors in video streams. The paper examines various techniques, including machine learning algorithms and deep learning models, that have been applied to video data for real-time anomaly detection. The authors identify key challenges such as handling high-dimensional data and the need for efficient processing to detect anomalies in dynamic environments. This work provides insights into how video analysis can be applied to security and surveillance systems, but also touches on related techniques for detecting anomalies in other forms of streaming data, including text and social media. The study is valuable for those working on real-time trend detection systems, as it offers a parallel perspective on the challenges of processing dynamic data streams and detecting emerging patterns.

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.
Gurcan et al. explore the use of probabilistic topic modeling techniques, such as Latent Dirichlet Allocation (LDA), to identify latent topics and trends in software engineering research over several decades. By analyzing a vast corpus of academic publications, the authors demonstrate how topic modeling can uncover hidden themes and reveal shifts in research interests over time. The paper highlights the importance of modeling temporal trends and the evolving nature of academic fields, offering a powerful tool for researchers interested in identifying emerging areas of interest within specific domains. The methodology proposed in this work is applicable to various fields beyond software engineering, including social media trend analysis and news topic detection, as it provides a robust framework for tracking the evolution of topics and trends over time. The study emphasizes the potential of topic modeling as an analytical tool for understanding long-term trends in academic and research contexts.

PROPOSED METHODOLOGY

This study presents a novel method for detecting and analyzing emerging topic trends in streaming text data using a Sequential Evolution Model (NSEM) that leverages distributed word representations. The process begins with the collection of data from a reliable news source, followed by chronological segmentation to track topic evolution over time. Preprocessing steps such as stopword removal, tokenization, and normalization are applied to clean the dataset and remove irrelevant noise. Word2Vec embeddings are then generated, using the Continuous Bag of Words (CBOW) approach to create separate models for each time segment, capturing semantic relationships effectively. Central to this methodology, the News Sequential Evolution Model (NSEM) analyzes the

evolution of word associations over time by comparing embeddings across different time periods. It detects emerging trends by identifying shifts in semantic similarity and pinpointing key influential words. A visual display model is incorporated to improve interpretability, allowing users to track changes in topic prominence, while a knowledge graph is used to visualize the relationships between significant terms, providing a deeper understanding of trend development. The methodology is implemented using Python and various NLP libraries, with experiments conducted on multi-year streaming news data. Hyperparameters, such as vector size and context window, are fine-tuned to optimize performance, and the approach is evaluated based on its consistency in detecting trends and its alignment with real-world events. This method outperforms traditional trend detection techniques by enabling real-time monitoring, utilizing chronological segmentation, and enhancing interpretability through visual and knowledge graph representations. By leveraging distributed word representations, this approach provides valuable insights into the dynamic nature of textual data, assisting researchers, businesses, and analysts in making informed decisions across a wide range of domains.

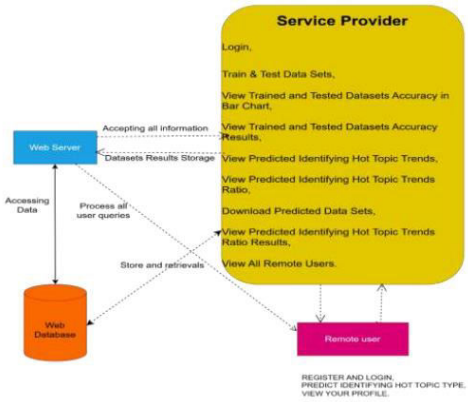

Fig: System Architecture

CONCLUSION

This study introduces a novel approach for detecting and analyzing emerging trends in streaming text data using a Sequential Evolution Model (NSEM) that leverages distributed word representations. By employing Word2Vec embeddings, this method tracks the evolution of word relationships over time, offering a more dynamic and insightful analysis of trending topics. Unlike traditional techniques, the proposed model segments the data chronologically, providing a clearer understanding of trend development across various time intervals. The integration of a visual display model and a knowledge graph enhances the interpretability of detected trends, allowing users to explore topic relationships in an intuitive manner. Experimental results demonstrate the model's effectiveness in identifying significant trends and providing insights that traditional methods often overlook. The practical applications of this research span areas such as market analysis, media monitoring, social trend prediction, and business intelligence, enabling decision-makers to adapt more efficiently to changes in public interest and industry dynamics. Future work could further enhance this approach by incorporating deep learning models, multimodal data, and real-time adaptive learning techniques to improve prediction accuracy. Additionally, refining the visualization tools could offer even deeper insights into complex topic relationships. Ultimately, this method contributes to advancing natural language processing and trend analysis, offering a scalable and efficient solution for understanding dynamic textual data streams.

REFERENCES

[1] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality,"

Advances in Neural Information Processing Systems, vol. 26, 2013.

[2] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in

Proc. 2013 Conf. North American Chapter of the Association for Computational Linguistics: Human

Language Technologies (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.

Supercomput., 2021.

[6] R. Raja, et al., "Analysis of anomaly detection in surveillance video: recent trends and future vision,"

Multimedia Tools Appl., 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. 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