IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 August 2023, accepted 31 August 2023, date of publication 7 September 2023, date of current version 14 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312764

RESEARCH ARTICLE

Identifying Hot Topic Trends in Streaming Text Data Using News Sequential Evolution Model Based on Distributed Representations

ZOHAIB AHMAD KHAN[1], YUANQING XIA[1], (Senior Member, IEEE), SHAHZAD ALI[1], JAVED ALI KHAN[2], S. S. ASKAR[3], MOHAMED ABOUHAWWASH[4,5], AND NORA EL-RASHIDY[6]

[1]School of Automation, Beijing Institute of Technology, Beijing 100081, China
[2]Department of Computer Science, Faculty of Physics, Engineering, and Computer Science, University of Hertfordshire, AL10 9AB Hatfield, U.K.
[3]Department of Statistics and Operation Research, College of Science, King Saud University, Riyadh 11451, Saudi Arabia
[4]Department of Computational Mathematics, Science, and Engineering (CMSE), Michigan State University, East Lansing, MI 48824, USA
[5]Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt
[6]Machine Learning and Information Retrieval Department, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafr El-Sheikh 6860404, Egypt

Corresponding author: Yuanqing Xia (xia_yuanqing@bit.edu.cn)

This project is funded by King Saud University, Riyadh, Saudi Arabia. Researchers Supporting Project number (RSP2023R167).

ABSTRACT Hot topic trends have become increasingly important in the era of social media, as these trends can spread rapidly through online platforms and significantly impact public discourse and behavior. As a result, the scope of distributed representations has expanded in machine learning and natural language processing. As these approaches can be used to effectively identify and analyze hot topic trends in large datasets. However, previous research has shown that analyzing sequential periods in data streams to detect hot topic trends can be challenging, particularly when dealing with large datasets. Moreover, existing methods often fail to accurately capture the semantic relationships between words over different time periods, limiting their effectiveness in trend prediction and relationship analysis. This paper aims to utilize a distributed representations approach to detect hot topic trends in streaming text data. For this purpose, we build a sequential evolution model for a streaming news website to identify hot topic trends in streaming text data. Additionally, we create a visual display model and knowledge graph to further enhance our proposed approach. To achieve this, we begin by collecting streaming news data from the web and dividing it chronologically into several datasets. In addition, word2vec models are built in different periods for each dataset. Finally, we compare the relationship of any target word in sequential word2vec models and analyze its evolutionary process. Experimental results show that the proposed method can detect hot topic trends and provide a graphical representation of any raw data that cannot be easily designed using traditional methods.

INDEX TERMS Topic trends, news sequential evolution model, stream text analysis, visual display model, knowledge graph, distributed representations.

I. INTRODUCTION

Detecting hot topic trends in real-time is critical in many fields, including marketing, technology, finance, and politics. However, traditional approaches to trend analysis often fall short when it comes to understanding complex and nuanced language use in a continuous stream of data. This is where distributed representation models, such as word2vec come in. Word2Vec allows grouping similar words together and implementing learning algorithms to improve performance on natural language processing tasks [1]. The model has attracted much attention due to its ability to construct the semantic context of words [2], [3]. It contains many algorithms and functions and can be implemented in Java, C, and Python. In short, word2vec is a tool used for computing the vector representation of words. It inputs value as text

The associate editor coordinating the review of this manuscript and approving it for publication was Porfirio Tramontana.

and gives output as word vectors. Although the usage of distributed representation models for creating embeddings is widespread, many unanswered questions remain about the factors that influence its results and its true capabilities [4], [5]. These models can efficiently capture the semantic and syntactic relationships between words and phrases, allowing for more accurate and precise trend analysis. In particular, the use of distributed representation models in a distributed computing environment can enable real-time processing of massive amounts of data, making it possible to detect and respond to emerging trends faster than ever before. Therefore, developing and applying distributed representation models for trend analysis is an area of growing importance and interest.

Some of the current issues in hot topic trend detection include the difficulty in handling large amounts of data, as well as the challenge of detecting subtle shifts in language use and topic evolution over different time spans. Different areas of application such as bioinformatics, data mining, speech recognition, remote sensing, multimedia, text detection, localization, and others, require different techniques to be utilized. Therefore, there is no single technique that can be applied universally across all these areas [6]. Understanding the trends in software engineering [7] emphasized the importance of deeper analysis and a systematic approach. In [8] the objective of the study was to examine the research trends in Science, Technology, Engineering, and Mathematics (STEM) education, while [9] gives importance to describing the fields of study and trends in computational thinking (CT), and [10] detected IoT Botnet in 5G Core Network. In [11] the author aimed to hidden discover topics and trends within historical incident reports of the Air Traffic Control (ATC) system. The research work of [12] introduces Bangla-BERT, a monolingual BERT model designed for the Bangla language. Additionally, many traditional methods for analyzing language and identifying topics rely on handcrafted features and rule-based approaches, which can be time-consuming and may not generalize well to different datasets. Furthermore, many of these methods may not be suitable for real-time processing, which can be important for applications such as social media monitoring and sentiment analysis. Therefore, there is a need for new, more efficient, and scalable methods for analyzing language use and detecting hot topic trends in data streams.

The field of natural language processing has experienced significant advancements in recent years, with the rise of machine learning techniques and the availability of large datasets. In previous times, users were required to communicate their requirements and intentions by composing a well-defined document in everyday language [13]. With the growth of social media platforms and online news websites, analyzing text data has become crucial in understanding public opinion, and consumer behavior, and predicting future trends. Streaming text data is becoming more prevalent, and traditional batch processing techniques are no longer sufficient to handle the high volume and variability of this data. One key challenge in this domain is identifying hot topic trends in real-time, as they emerge and evolve. Traditional methods for trend detection often rely on handcrafted features or require significant human intervention, making them slow and prone to errors. In contrast, machine learning techniques based on distributed representations have shown great promise in automatically learning patterns and relationships in large volumes of textual data. Therefore, there is a need for simultaneous analysis of streaming text data to capture evolving trends and patterns.

In this paper, we aim to address this need by proposing a novel approach for detecting hot topic trends in streaming news data using a sequential evolution model based on distributed representations. The approach has the potential to provide valuable insights to market analysts, news agencies, and researchers in various fields. The ability to analyze trends in large volumes of text data is important for a wide range of applications, including trend detection in social media and news analysis. However, existing methods for trend analysis are often limited by the lack of robust techniques for analyzing the relationships between target words. To address this problem, we propose a novel NSEM for analyzing the difference in sequential periods between different data sets. Our approach uses word2vec models to analyze the semantic relationships between words for each dataset separately and identifies trends by comparing these models over periods. This approach allows us to accurately identify trends, provide a data visualization model, and create a knowledge graph of raw streaming news data.

A. MOTIVATION

The exponential growth of online news and social media platforms has led to an information explosion, resulting in an overwhelming amount of data. As a consequence, it has become increasingly difficult to identify and analyze the most relevant and trending topics over different time periods. Our motivation for this research stems from the pressing need to develop efficient methodologies that can accurately detect and track hot topics from vast and dynamic datasets. By gaining a comprehensive understanding of these trending topics and their temporal patterns, we aim to offer valuable insights that can benefit various industries and academic disciplines. Through our proposed approaches and models, we aspire to unlock the potential of large-scale streaming news data and provide actionable information for media professionals, business analysts, and researchers to make informed decisions and gain deeper insights into the changing trends of the digital world.

B. CONTRIBUTIONS

The paper presents several contributions toward trend detection and analysis. Firstly, we explore the use of distributed representations as a promising approach for trend detection, which allows for comprehensive analysis of data by capturing semantic relationships between topics. Moreover, we propose the News Sequential Evolution Model (NSEM) which uses distributed representations to analyze trend relationships between words in large datasets. Additionally, we introduce a unique chronological analysis of trends and relationships between words in different datasets, which considers how they change during different time periods for a more accurate understanding of their evolution. Furthermore, our visualization display model leverages data visualization theory to identify hot topic trends across various domains, providing valuable insights for decision-making and staying ahead of the competition. Finally, we present a word2vec-based knowledge graph representation technique that contributes to identifying patterns and relationships in streaming news data, enabling applications such as question-answering systems and recommendation engines. To the best of our knowledge, this paper is the primary work in identifying topic trends from stream text news based on distributed representations. These contributions could potentially have significant implications for a range of applications in the field.

C. PAPER ORGANIZATION

The remaining paper is organized into four sections. In Section II, related work is reviewed, recent advances in the field are highlighted and our proposed method is compared to state-of-the-art approaches in the literature. Section III provides an in-depth explanation of the concepts behind NSEM and an overview of our proposed method. The experimental results that support the superiority of our method over existing approaches are presented in Section IV. Finally, Section V concludes the paper, summarizing the main findings and suggesting potential areas for future research. The paper presents an innovative and effective approach for analyzing trends, which has been shown to outperform existing methods in the literature.

II. RELATED WORK

Since Mikolov et al. [1], [2], [3] introduced the word2vec model, many studies have applied the model to learn word embeddings. However, this work focuses on a different application of the word2vec and implements the model to analyze the trend relationships between target words. Additionally, our work introduces a novel model called NSEM to analyze the difference in sequential periods between different data sets, a visual display model, and a knowledge graph to represent trends. Thus, our work is introducing new contributions and applications in the field. Dynamic topic modeling was presented by [14], where a family of probabilistic time series models is developed to analyze the temporal evolution of topics within extensive document collections. The proposed approach involves utilizing state space models on the natural parameters of multinomial distributions that represent the topics. However, our work differs from dynamic topic modeling in several key aspects. While both approaches aim to capture the temporal dynamics of the topics, they employ different methodologies and have distinct strengths.

Dynamic topic modeling (DTM) is a probabilistic modeling technique that enables the discovery of topics that evolve over time in a given document collection. It considers the temporal ordering of documents and captures the changing prevalence of topics over time. DTM can be effective in identifying topic shifts, tracking topic evolution, and revealing latent thematic patterns in textual data. On the other hand, our proposed NSEM approach focuses on the detection of hot topic trends specifically. We leverage the distributed representation approach, such as word2vec, to capture semantic relationships and patterns within the data. This enables us to identify and analyze trends that gain prominence or decline during diverse time periods. It is important to note that dynamic topic modeling can be a valuable technique in various applications, particularly when the objective is to analyze the evolution of topics over time at a more granular level. In contrast, our focus is on identifying and understanding the trends that are currently popular or gaining traction within a given dataset. Overall, while both dynamic topic modeling and our proposed approach aim to capture temporal dynamics, they differ in their methodologies, objectives, and specific contributions to the field of trend detection. Each approach has its strengths and can be applied depending on the specific research goals and requirements.

A. MACHINE LEARNING-BASED APPROACHES

Word2Vec model was applied to hot topic trend detection in [4]. However, they did not consider the utilization of a word2vec-based knowledge graph, and their findings were not supported by a thorough comparison with existing methods. It is worth noting that in [15], a word2vec-based model was proposed specifically for identifying COVID-19 viral sequences using protein vectors. The researchers applied five different machine learning models for classifying COVID-19 viral sequences. The authors of [16] automatically interpreted Twitter topic trends and described two factors that influence social media topics. In the research done by [17], the authors provide a comprehensive analysis of software testing topics and trends over the past 40 years, while our work typically focuses on popular or trendy topics from stream news that are currently generating attention. The study's methodology involves probabilistic topic modeling of a large corpus of published articles, while we have used distributed representations which is more accurate. Overall, the study provides a more in-depth and historical perspective on software testing topics and trends, while hot topic trends tend to reflect more immediate and current interests.

According to [18] their work analyzes customer reviews to identify topics and trends, as hot topic trends typically refer to current popular or frequently discussed topics. The two concepts are related, this paper specifically focuses on customer reviews and LDA topic modeling, whereas we utilize word2vec models. Additionally, the authors of the paper were unaware of NSEM, visualization models, and knowledge graph representations of the trends. According to the research in [19] their novel approach uses a weighted temporal feature to cluster finance journal abstracts from 1974 to 2020. They detect emerging trends that are missed by standard clustering methods, using a metric that combines silhouette scores and cluster standard deviation over time. They validate the results with expert judgment and suggest that their method can be applied to other fields. However, they have not mentioned chronological analysis, visual model, or word2vec-based knowledge graph. Our work will be the first to include these approaches, we will not only analyze the relationships between words but also take into account how those relationships change over a period of time.

According to the contribution of [20], a semi-supervised approach was proposed for detecting anomalies in multivariate categorical data. The approach learns the probability distribution of normal instances using a Gaussian process generative model. This non-parametric Bayesian model adapts its complexity to the data size, making it effective for small training datasets. Comprehensive experiments show the effectiveness of this approach. However, their method is only effective for small datasets. In our case, we are dealing with a continuous stream of news coming from the internet, and we use the word2vec model to detect hot topic trends, which is more suitable for large datasets. In [21] the author provides a clear overview of the research topic, methodology, and findings. However, it lacks details on the specific techniques used for content analysis, sentiment analysis, and Latent Dirichlet Allocation, which could limit the reproducibility and generalizability of the study. Additionally, the paper does not provide much context on why identifying the most-discussed topics for different countries is important or how the findings could be applied in practice. In contrast, our work provides a more comprehensive overview of the research approach and highlights the significance of the research.

B. DEEP LEARNING-BASED APPROACHES

Based on the research work of [22], the author proposes deep neural architectures that use Long Short-Term Memory (LSTM) to predict trending research topics by deep neural network-based content analysis. The models fall into two groups: one considers each paper as a sequence of keywords, while the other represents each paper as a vector. The models are evaluated on academic paper collections from the Microsoft Academic Graph data set. The proposed method in the paper is designed to predict the future use of a specific keyword based on a collection of publications. The paper's proposed method is focused on predicting the future use of a specific keyword, while hot topic trends are typically used to identify popular topics that people are currently interested in.

As described in [23], the authors propose a novel approach for predicting bitcoin trends using the One-Dimensional Convolutional Neural Network (1D CNN). They develop a methodology for building datasets that include social media data, blockchain transaction history, and financial indicators, and also introduce a cloud-based system with a distributed architecture for collecting data. The results show that their 1D CNN model outperforms LSTM models in predicting bitcoin trends. In comparison, our paper presents a unique methodology for analyzing streaming news data by building NSEM using word2vec models. This approach is not commonly used in the field of data analysis and has wide-ranging applications in areas such as marketing, sentiment analysis, and finance. Our work is also well-written and accessible, even to readers who are not familiar with the technical jargon used in the field of data analysis.

Comparing our work to [24], both research papers propose novel approaches to solving problems in the field of data analysis. Their work presents a fuzzy detection system for rumors using an explainable adaptive learning method that utilizes a Graph-GAN model with fine-grained feature spaces and continuous adversarial training. Our work, on the other hand, explores the use of distributed representations in streaming data clustering and develops a sequential evolution model for a streaming news website with the insights of NSEM, a word2vec-based knowledge graph, chronological order of the data, and a visualization model. A significant advantage of the proposed method is its ability to identify trends in the relationship between target words that cannot be easily measured through traditional methods.

According to the authors of [25], their work proposes a deep learning approach for detecting and classifying malware in healthcare cyber-physical systems. The system uses byte sequences from ELF files and automatically extracts features while identifying CPU architecture. The proposed method is robust and generalizable for malware detection and classification, and CPU architecture identification in healthcare cyber-physical systems. In comparison to the this work, our proposed method offers a wide range of novelty. We introduce a new approach specifically designed for analyzing the evolution of stream news data, which can be of great interest to researchers working in the fields of linguistics, trend analysis, and media discourse. Our work is versatile as their work focuses specifically on malware detection in healthcare systems, while our method could potentially be applied to a wide range of streaming data, such as social media, healthcare systems, financial markets, or scientific publications. Our work is Interpretable as word2vec models are based on distributed representations of words, which can be more interpretable than the deep learning models used in their work. This may be particularly useful in settings where it is important to understand how the model is making predictions or identifying patterns in the data. Furthermore, our work is flexible as it allows the analysis of multiple target words, rather than just malware detection or CPU architecture identification as in their work. This flexibility could make the method more adaptable to different research questions or domains.

C. ADVANCEMENTS IN DEEP LEARNING MODELS FOR DETECTING HOT TOPIC TRENDS: FROM WORD2VEC TO GPT-3 AND BEYOND

Recent advances in deep learning have revolutionized the field of natural language processing and have been extensively utilized in detecting hot topic trends. From the early success of word2vec to state-of-the-art models like

GPT-3, a plethora of deep learning techniques have emerged for effective topic detection and classification. These advancements include models such as BERT, Transformer-based architectures, transfer learning, pre-trained models, deep learning architectures for text classification, XLNet, RoBERTa, ELECTRA, T5, ALBERT, BART, and XLM-ROBERTa, etc. These models have demonstrated exceptional performance by leveraging large-scale language models, transfer-based learning, and innovative architectures, allowing for a more accurate and nuanced analysis of hot topic detection. The continuous development of deep learning methods in this domain holds great promise for uncovering meaningful insights and trends in large text collections, enabling researchers and practitioners to stay at the forefront of emerging topics.

Recent work by [26] has explored the detection and tracking of hot topics and trending social media events in real-time social networks. Their approach incorporates BERT and memory graph, combining Transformers with an incremental community detection algorithm to capture semantic relations and improve topic analysis. Additionally, they leverage named entity recognition from multimodal data and utilize NoSQL technologies to enhance system performance. The study of [27] applies the unsupervised BERT model for sentiment classification and the TF-IDF model for topic summarization. The findings reveal four major concerns related to the virus: origin, symptoms, production activity, and public health control. Transparent information sharing and scientific guidance are identified as potential ways to address public concerns effectively. The research project of [28] introduces Chat2VIS, a novel system that utilizes large language models (LLMs) like ChatGPT and GPT-3 to convert natural language into code for generating visualizations.

The research work of [29] utilizes recent Transfer Learning models to detect clickbaits on Twitter. By adapting models like BERT, XLNet, and RoBERTa with novel configuration changes, including model expansion and data augmentation, the authors achieve improved accuracy compared to a benchmark model. The experiments cover different scenarios and fine-tuning approaches, with RoBERTa integrated with an additional non-linear layer showing the highest performance. Furthermore, the work of [30] introduces two Bidirectional Long Short-Term Memory (BiLSTM) architectures with sentence transformers for fake news detection. The proposed models address both mono-lingual and cross-lingual fake news detection tasks, aiming to mitigate the harmful effects of fake news in society. The study explores the potential of large language models, specifically GPT-3, in identifying and classifying sexist and racist text [31]. Their findings suggest that large language models have the capability to contribute to hate speech detection and, with continued advancements, may be utilized in combating hate speech.

The field of Deep Learning (DL) has seen remarkable progress [32], but the increasing complexity of DL models raises concerns about feasibility. Efficient DL algorithms, like Spiking Neural Networks (SNNs), offer the potential to reduce power consumption without sacrificing performance. However, training SNNs is challenging due to non-differentiable activation functions. Despite this, researchers have made efforts to develop competitive SNN models, focusing on biologically inspired strategies. This survey explores SNN concepts, recent trends in learning rules, and network architectures, and provides practical considerations and research opportunities [32]. In the study conducted by [33], a baseline for detecting misleading reviews was established using two widely used language models: BERT and ELECTRA. The study conducted in [34] investigates the performance of the T5 architecture in comparison to three other state-of-the-art models across multiple tasks and datasets. By augmenting the training data and utilizing explainable AI techniques, the researchers achieved near-state-of-the-art results on selected tasks.

While numerous interesting and meaningful works have been presented in the field, our work specifically focuses on the detection of topic trends. We aim to examine the relationship between target words using distributed representations. We introduce a new model called NSEM to analyze the differences in sequential periods between various datasets. Multiple word2vec models are constructed for each dataset separately, and the datasets are then segmented in sequential order for comparison and analysis. This approach gives precise results in the form of trends that can be utilized to construct a visual display model and generate a knowledge graph of raw stream data from a news website using text similarity via word2vec. Additionally, the study uncovers significant information about NSEM.

III. METHODOLOGY OVERVIEW

Enhancing our previous work [4], we will explain the methodology used for detecting hot topic trends and describe the research data set utilized for eliciting requirements.

Models help provide learners with representations of scientific concepts, aiding them in understanding complex ideas. However, users must establish a clear connection between the model and the reality it represents. This process involves evaluating both the model itself and its relationship to the scientific concept being studied [35]. The swift increase in the adoption of the Internet, mobile, and social media apps has made it increasingly feasible to create channels that connect a vast number of users [36]. The nature of social media, particularly Twitter, where everyone can instantly post, share, and gather information related to topic trends, presents a fertile ground where exact information can be hard to find [37]. Different models have their specific rules and limitations regarding when and where they can be applied. Thus, a model without shortcomings is required to analyze streaming news data considering the sequence period, based on future needs. The NSEM is first used in our approach, making our work more useful to upcoming researchers and obtaining more accurate results. Without NSEM, it is challenging to analyze the difference between sequential periods, especially in streaming news datasets. In these datasets, the data stream comes from the web, and users cannot get interim information about the data.

A. NEWS SEQUENTIAL EVOLUTION MODEL

"News Sequential evolution model is a novel model based on distributed representations to analyze the difference of sequential period word embedding models from sliced streaming data, which aim to find the evolutionary process of correlations in each model."

B. INTRODUCTION

The NSEM is not only useful for analyzing sequential periods in streaming news data but it can also be applied to other types of data streams. Such as social media updates, financial market data, and sensor readings. Its flexibility and adaptability make it a valuable tool for data analysts and researchers in various fields. Moreover, the NSEM model has several advantages over other models, such as its ability to handle missing data and its capability to capture the nonlinear relationships between variables. Additionally, it can accommodate different types of data making it suitable for analyzing diverse data streams. In conclusion, the NSEM model is a powerful tool for analyzing sequential periods in data streams, enabling researchers and analysts to obtain more accurate results and insights. Its versatility and flexibility makes it a valuable asset for various fields, and its unique features make it a superior option to other models. In this paper, we present an in-depth analysis of NSEM, including its features, applications, and benefits, and demonstrate its effectiveness in various fields.

C. WORKFLOW DESIGN AND EXECUTION

The proposed method is visually represented in Fig. 1. We collected a continuous stream of news data from the internet and segmented it into separate datasets for each year from 2010 to 2016 (excluding 2014). We then applied the word2vec NSEM model to each dataset segment, analyzed the data, selected topics, and compared sequential models to obtain trend results. Many technology users, market analysts, researchers, and financial experts seek to find trends in their products and competitors online. They continuously monitor the actions of their product or competitors and follow each step they take. Some researchers study trend topics and use various techniques to solve the problem. Some use offline data to find trends in their favorite topics, while others use text, image, and video data from social media to try to address topic trend demands. Numerous models can solve the problem of topic trends and related issues, but some models may be difficult to apply, while others may have shortcomings. In the case of streaming news users may experience data order issues. Researchers must obtain information about topic cluster size, the number of documents containing topic keywords, topic term weights, and more for experimental measurements. Topic significance is usually measured on a time scale, such as time, date, week, month, or year. In this work, we assess the significance of topics on a larger time scale, specifically on a yearly basis.

In a nutshell, if the data is streamed, it is difficult for the user to know the time, date, month, or year when the data was collected, as well as topic term weights, topic cluster sizes, etc. If the users use the NSEM model to find the evolution of the correlations of each data set, they will be able to easily obtain all relevant information about the data.

D. STEP-BY-STEP PROCESS OF THE NSEM MODEL

Here is a step-by-step process of how NSEM works.

1) WORD2VEC NSEM MODELS

The first step involves building word2vec-based NSEM models using the streaming data collected from the internet. The data will then be segmented into different portions, depending on the structure and function of the experiment. This segmentation will help ensure that the models are trained on relevant data and can accurately capture the semantic relationships between words.

2) DATA ANALYZING

In the next step, we examined and interpreted the collected data to prepare it for experimentation. As for word2vec, it is important to have a dataset with a large amount of text data. This is because the performance of word2vec improves with larger datasets. Additionally, the dataset should be preprocessed to remove any unnecessary information, such as stop words and punctuation and the data should also be tokenized. We evaluated the accuracy and effectiveness of different models and compared their performance. We also assessed the quality of the data used to train the models. Overall, our goal in analyzing the data is to extract useful insights that can inform our research or decision-making

3) TOPIC SELECTION

In step 3 we selected three topics "LG", "Apple", and "China", using our domain knowledge. Ultimately, selecting these topics aims to identify an area that is both relevant to our research and feasible to explore, interesting for users and readers, while also allowing for potential insights and contributions to the field.

4) COMPARE THE SEQUENTIAL PERIOD OF EACH MODEL

In the final step, we compare the sequential model of six years of NSEM models using the Continuous bag-of-word (CBOW) word2vec model with a multi-word neural network content setting. To facilitate a better understanding of our research methodology, we have included a visual representation of an artificial neural network [38] in our manuscript. This illustration serves as a fundamental building block of a neural network and aids readers. The purpose is to enable a better understanding of our research methodology, particularly for readers who may be less familiar with neural networks and their components. The process is followed by an explanation of why and how the CBOW model is utilized in our proposed approach, following the approach of [38].

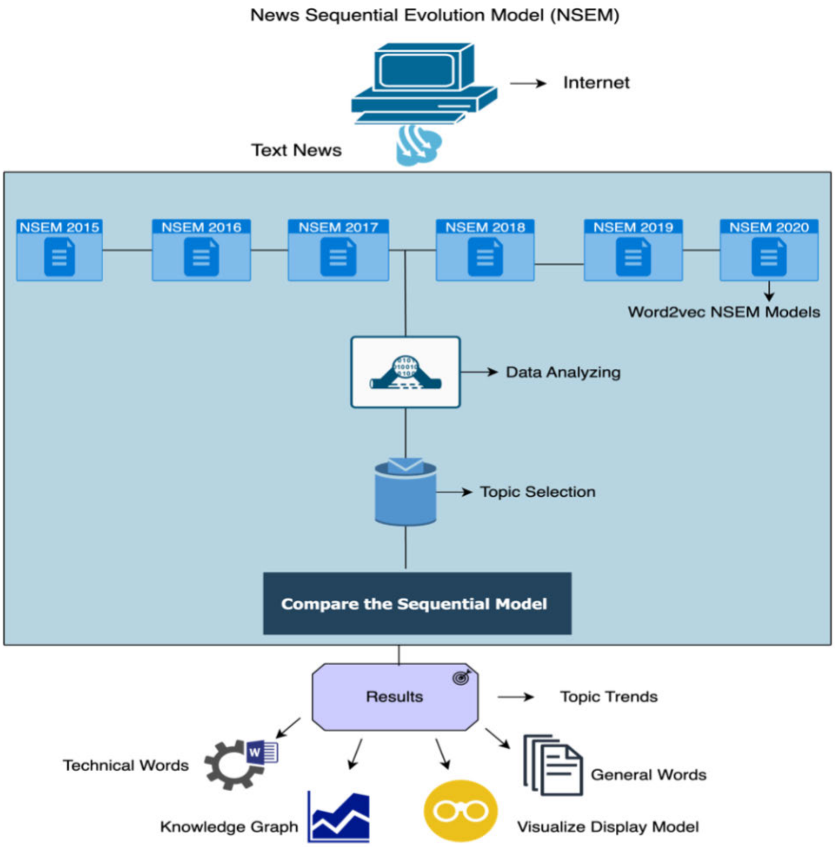

FIGURE 1. Workflow of news sequential evolution model.

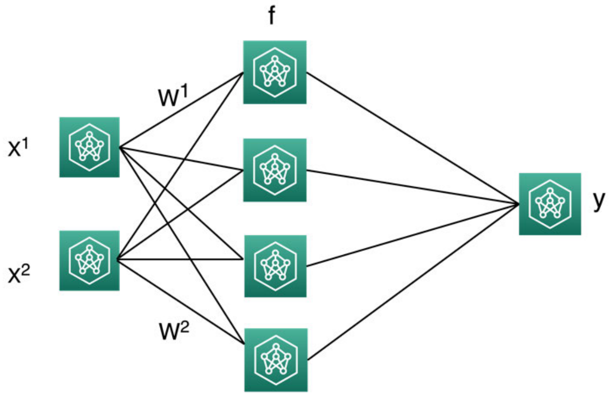

FIGURE 2. Artificial neural network.

Figure 2 shows the (unit) of an artificial neural network, $\{x^1, \cdots, x^k\}$ are key in values; weights are from $\{\mathrm{w}^1, \ldots, \mathrm{w}^k\}$; y is the output production of the scalar; and f is the sum point (also called transfer/ decision/activation point).

The choice of the Continuous Bag of Words (CBOW) model was deliberate, considering its effectiveness in word embedding and semantic similarity tasks. While it may be considered a relatively simpler approach compared to more complex models, such as DNN or ANN, it has demonstrated remarkable performance in various NLP applications, including trend detection and topic modeling. One of the main reasons for using the CBOW model is its efficiency in capturing semantic relationships between words within our given dataset. By focusing on the essential aspects of the model, we were able to achieve accurate results in identifying topic trends and correlations over time spans. Considering the nature of our research and the specific goals we aimed to accomplish, the simplicity of the CBOW model was advantageous, as it allowed us to efficiently analyze large-scale streaming news data. More complex models might introduce unnecessary computational overhead without significantly improving the outcomes. The CBOW model is a crucial component of our approach. A graphical representation of the Word2Vec CBOW model is provided in Figure. 3, illustrating its role in detecting semantic similarity and contributing to the identification of hot topic trends.

V × N matrix W represents the weights among the input layer with the hidden layer. $\mathrm{W}V \times N = \{wki\}$ and $\mathrm{W}'N \times V = \{w\text{'}ij$, The CBOW model takes the average of the background words instead of a straight repetition of the key in the vector of the input context and makes use of the result of

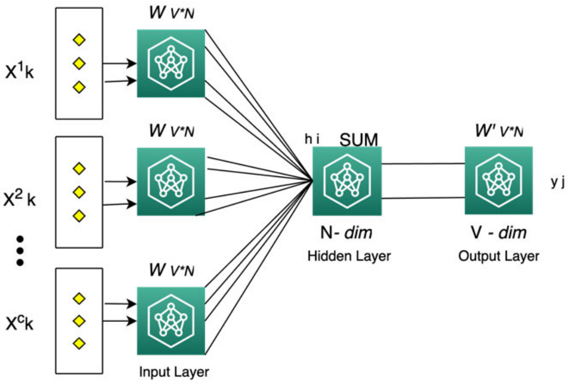

FIGURE 3. Continuous bag-of-word model.

the input hidden weight matrix and the average vector as the production during the computation of the hidden layer output.

$$\mathrm{h} = \frac{1}{c}W^T\left(x_1 + x_2 + \cdots + x_c\right) \quad (1)$$

$$= \frac{1}{c}(v_{w_1} + v_{w_2} + \cdots + v_{w_c}) \quad (2)$$

Where the number of words in the context is C, the words in the context are from $w_1, \ldots, w_c$, and the key in of a word w is v_w. The loss function is expressed as follows:

$$E = -logp(w_o|w_{1,}w_{2,}\cdots, w_i; C) \quad (3)$$

$$= -u_j^* + log\sum\nolimits_{j'=1}^{v} exp(uj') \quad (4)$$

$$= -v'^T_{wo}\cdot h + log\sum\nolimits_{j'=1}^{v} exp(v'^T_{wj}\cdot h) \quad (5)$$

The hidden output weight is expressed as:

$$V'^{(new)}_{wj} = V'^{(old)}_{wj} - \eta.e_j.h,\ for\ j = 1, 2, \ldots, V. \quad (6)$$

where e_j is a prediction error. We must apply the above to each element of the hidden-output weight matrix for every training instance.

The equation for input-hidden weights is:

$$V^{(new)}_{WI,c} = V^{(old)}_{WI,c} - \frac{1}{c}.\eta.EH^T,\ for\ c = 1, 2, \ldots, C. \quad (7)$$

$V_{WI,c}$ is the key in vector in the input context of the word c-th; is a positive learning rate; and EH $=\frac{\partial E}{\partial h_i}$.

E. INNOVATIONS BY NSEM

In recent years, the use of machine learning models has become increasingly popular for studying large collections of text data. One of the most important tasks in this field is identifying sequential patterns in the data, which can help users understand the evolution of topics over time. To address this challenge, many academics have developed different models to target the analysis of streaming news data. One of the advantages of NSEM is its ability to find the sequential periods of the data without any time constraints, making it ideal for studying large collections of constantly-updated text data. In addition to identifying sequential patterns, NSEM can also identify the main topics in the data, their importance, and correlations. This is achieved through a combination of unsupervised and supervised learning techniques that enable the model to learn from the data and make accurate predictions. Once the main topics and their correlations have been identified, users can easily understand the periods of data and conduct experiments to gain further insights. For example, they can analyze how the importance of different topics changes over different time spans or how the correlation between topics evolves over different time periods.

IV. EXPERIMENTATION

In this part, we use the proposed method to analyze the six datasets of streaming news data collected from the "businessinsider.com" website. We apply the NSEM model to these datasets and compare the results to assess the accuracy and effectiveness of the model. Specifically, we examine the output of the NSEM models to identify the main topics that exist in the data and their correlated topics. We also analyze how the importance of different topics changes over periods, and how the correlation between topics evolves over different time spans. Moreover, we compare the performance of different models, such as the continuous bag-of-words (CBOW) word2vec model with a multi-word neural network content setting, to see which model provides better results. We also assess the quality of the data used to train the models and make sure that it is proper for the experiments. By creating a knowledge graph and using a visualization display model, researchers can better understand the relationships between different topics and how they evolve. This can lead to insights and discoveries that might not be immediately apparent through traditional methods of data analysis.

The goal is to extract useful insights from the data that can inform future research or decision-making. By conducting experiments and analyzing the results, we can gain a deeper understanding of the topics and trends that exist in the data, and use this knowledge to make informed decisions or improve our future research.

A. DATASET

The dataset used for this research was collected from "businessinsider.com," a prominent financial news website headquartered in the USA. The website is owned by a German publishing house and operates international versions in the UK, China, Italy, Indonesia, Malaysia, Australia, and India. Additionally, it offers local versions in the form of German and Polish-language news sites. "Businessinsider.com" provides a wide range of valuable content, including quick, reliable, and insightful commentaries on various business verticals such as technology, finance, strategy, and politics from around the world. The website serves as a comprehensive source for the latest information and breaking news alerts related to these industries. Technology enthusiasts and business professionals find the site particularly relevant as it offers substantial and well-researched news coverage [4].

Despite the availability of a substantial amount of high-value streaming news on various websites, we lack appropriate tools to track the hotspots and their evolution over different time spans. To address this, a stream of news articles from the website was collected and six datasets were compiled, each containing text data from a specific year within the timeframe of 2010-2016 (except 2014). After continuously receiving the data from the streaming source it was stored into buffers, where the data underwent preprocessing before being converted to.txt files for analysis to extract trends and patterns using Word2Vec NSEM models. We recognize the importance of using this rich and diverse dataset from ''businessinsider.com'' for our research, as it allows us to explore trends and correlations in a variety of fields and domains. These datasets were created to capture the evolving trends and topics over a time span, allowing us to analyze the changes and patterns in hot topic detection. Various data processing and general analysis steps were performed, to prepare the data for experimentation.

B. EXPERIMENTAL SETUP

To conduct the experiments described in this article, we used an Intel(R) Core (TM) i5 CPU and 8GB of RAM. We used PyCharm as our integrated development environment and the Python 2.10 interpreter for coding. The Word2Vec parameters are set as follows: 'vector_size' = 200, 'window' = 10 for ''LG'' and ''Apple'' and 20 for ''China'', and 'min_count' = 1. We employed a hold-out sample selection technique to thoroughly evaluate the accuracy and effectiveness of our model. Time parameters are not assumed in the proposed model, as we acknowledge that the consideration of time can be crucial in certain scenarios, particularly when analyzing fine-grained temporal trends at hourly, daily, or weekly intervals. However, in our study, we specifically focus on detecting hot topic trends on a yearly basis using data spanning from 2010 to 2016 (excluding 2014). In this context, our research methodology aims to identify overarching trends and patterns that evolve over longer time spans rather than pinpoint specific temporal occurrences. By analyzing yearly data, our objective is to provide insights into the broader trends and topics that gain prominence or decline over long time periods. We believe that this approach aligns with the specific nature of our dataset and research objectives. While incorporating time parameters could be a valuable avenue for future research, we argue that it is not necessary for achieving our current research goals. However, the methodology and framework presented in our work can be adapted to datasets with more precise time and date information, allowing for more detailed temporal analysis when available.

C. FEATURE SELECTION

Our analysis focuses on three specific topics: ''LG'' and ''Apple'' for technical trend detection, and ''China'' for general trend detection. These topics were chosen based on our domain knowledge and expertise in the subject area. For each year of the news data set, we use word2vec models to detect topic trends for each of these three topics. Specifically, we build word2vec models for different periods within each data set. We then identify the hot topic trends for each year of data for the selected topics and calculate the percentage of relationships among them.

D. RESULTS

In the results section, we aim to present the outcomes of our proposed method for identifying hot topic trends in text data streams using NSEM based on distributed representations. Our approach has the distinct advantage of offering trend analysis exclusively for technical words like ''Apple'' and ''LG'' as well as for general words such as ''China'', which can relate to various aspects of life. This unique feature of our proposed method sets it apart from previous trend detection methods. As far as we know, no other method has demonstrated such superiority in identifying trends and establishing their relationship percentage with a given word. Our results show the effectiveness of our proposed method and how it outperforms existing methods in detecting and analyzing trends in text data streams. Traditional performance metrics like accuracy or precision may not directly apply to word2vec semantic similarity, as it is based on vector representations of words rather than numerical percentages. We have obtained valuable insights from our experiments that have practical implications for various professionals. Furthermore, we focused on using qualitative assessments from domain experts to validate the semantic relationships between words and verify the quality of the word2vec model.

The paper proposes a methodology and toolbox for analyzing streaming data sources, providing users with insights into trending topics related to their products. Websites like ''businessinsider.com'' share important news from around the world that technology users and fans need to stay up-to-date on. Our toolbox allows users to precisely target the evolution of their favorite products, such as ''Apple'', technologies like ''VR'', and companies like ''Google'', and ''OpenAI''. It enables them to visualize results and see their relationship percentages for different time spans. Using this toolbox, users can easily track the evolution of their favorite products, technologies, and companies over time, making data-driven decisions. For example, businesses can understand how their products are perceived by the public and how their competitors are performing in the market. Technology enthusiasts can keep up with the latest developments and trends related to their favorite technologies. Visualizing results and relationship percentages for different periods can provide valuable insights into data trends and patterns. Ultimately, our proposed method offers a powerful solution for analyzing streaming data sources, empowering users to stay up-to-date with the latest trends and make informed decisions.

For the first experiment, we focused on analyzing the trends related to the word ''LG'' in each dataset. Our proposed method identified the top 10 distinct trends related to ''LG'' that were present within the given datasets. By visualizing these trends in a graphical format, we provided a clear

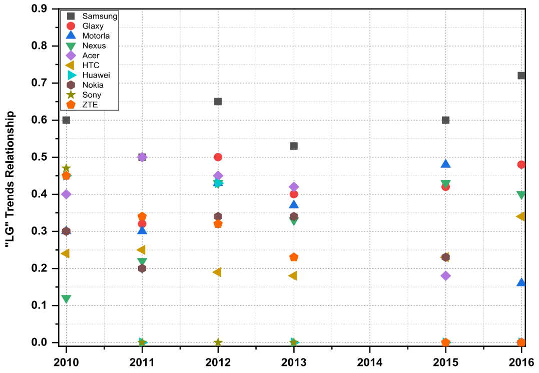

FIGURE 4. "LG" trends in each year of data.

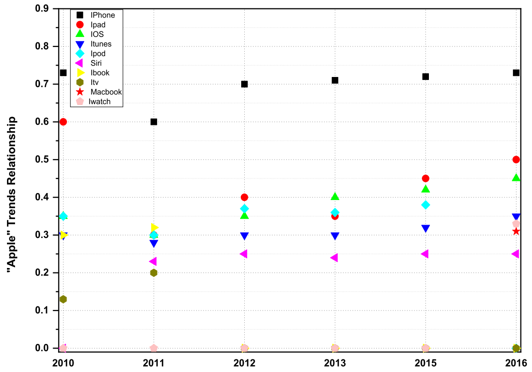

FIGURE 5. "Apple" trends in each year of data.

and comprehensive overview of the significant trends and their relationship changes over periods associated with the word "LG".

Figure. 4 illustrates how our proposed method effectively filters out low-value trends that do not significantly contribute to our suggested topic within the given dataset. For instance, the relationship between "LG" and its trends is only visible in the graph for some years when the relationship percentage is low. As the relationship percentage grows, so does the value of their relationship. However, our system automatically removes these low-value trends to ensure that the suggested topics are relevant and significant. Furthermore, Fig. 4 also demonstrates how the occurrence of appearance of a particular word in the datasets affects its relationship percentage with other words. For example, "Samsung" occupies the highest space in the graph because it appears in every dataset, and the relationship value between "Samsung" and "LG" is higher. On the other hand, the trend "Sony" only appeared in 2010 and is absent in the remaining datasets, resulting in a lower correlation percentage between "LG" and "Sony," and a smaller value. These findings demonstrate the robustness and effectiveness of our proposed method for trend detection in text data streams.

In the next experiment, we applied the same technique to the word "Apple" and obtained similar results. The top ten trends were then visualized in a graphical format to show their relationship percentage with the word "Apple" over these years. As seen in Fig. 5, the graph reflects the ups and downs of the identified trends related to "Apple", providing valuable insights into the changing relationships of the trends. This experiment reinforces the effectiveness and versatility of our proposed method in identifying and tracking hot topic trends in data streams.

As shown in Fig. 5, we used a similar technique as in Fig. 4 and obtained similar results. The trends that had small relationship values with "Apple" only appeared in one or two sets of data, whereas the words with larger distance values emerged in all datasets, and their relationship percentages were higher than others. In Fig. 5, "iPhone" became well-known in every set of data and stayed at the top of the graph since it is the nearest word to "Apple" in all the datasets. On the other hand, "IWatch" and "MacBook" only appeared in the dataset of 2016. This suggests that the relationship between "Apple" and these two trends was valued near or at zero before 2016.

Our proposed method has shown promising results in identifying relevant and significant trends for technical words. The analysis of the trends related to "LG" and "Apple" demonstrates how our method effectively filters out low-value trends and captures the relationship between different words over different time periods. The results also highlight that words that appear more frequently may have a stronger relationship with the given words, than words that appear less frequently. These findings demonstrate the strength and effectiveness of our proposed method for trend detection in text data streams. words. The analysis of the trends related to "LG" and "Apple" demonstrates how our method effectively filters out low-value trends and captures the relationship between different words over different time periods. The results also highlight that words that appear more frequently may have a stronger relationship with the given words, than words that appear less frequently. These findings demonstrate the strength and effectiveness of our proposed method for trend detection in text data streams.

E. VISUAL DISPLAY MODEL

One of the primary motivations for creating a visual display model in this study is to enhance the interpretability and communicability of our findings. While textual descriptions and statistical analyses are useful for conveying information, they can often be dense and difficult to comprehend for non-experts. By using a visual model, we can provide a clearer and more engaging representation of our results, which can help to bridge the gap between technical terminology and real-world applications. Additionally, a visual model can allow us to explore and discover new patterns and relationships in our data that may not be immediately apparent through textual or numerical summaries alone.

Figure. 6 demonstrates that our method can identify trends not only in one field but across different related and matching

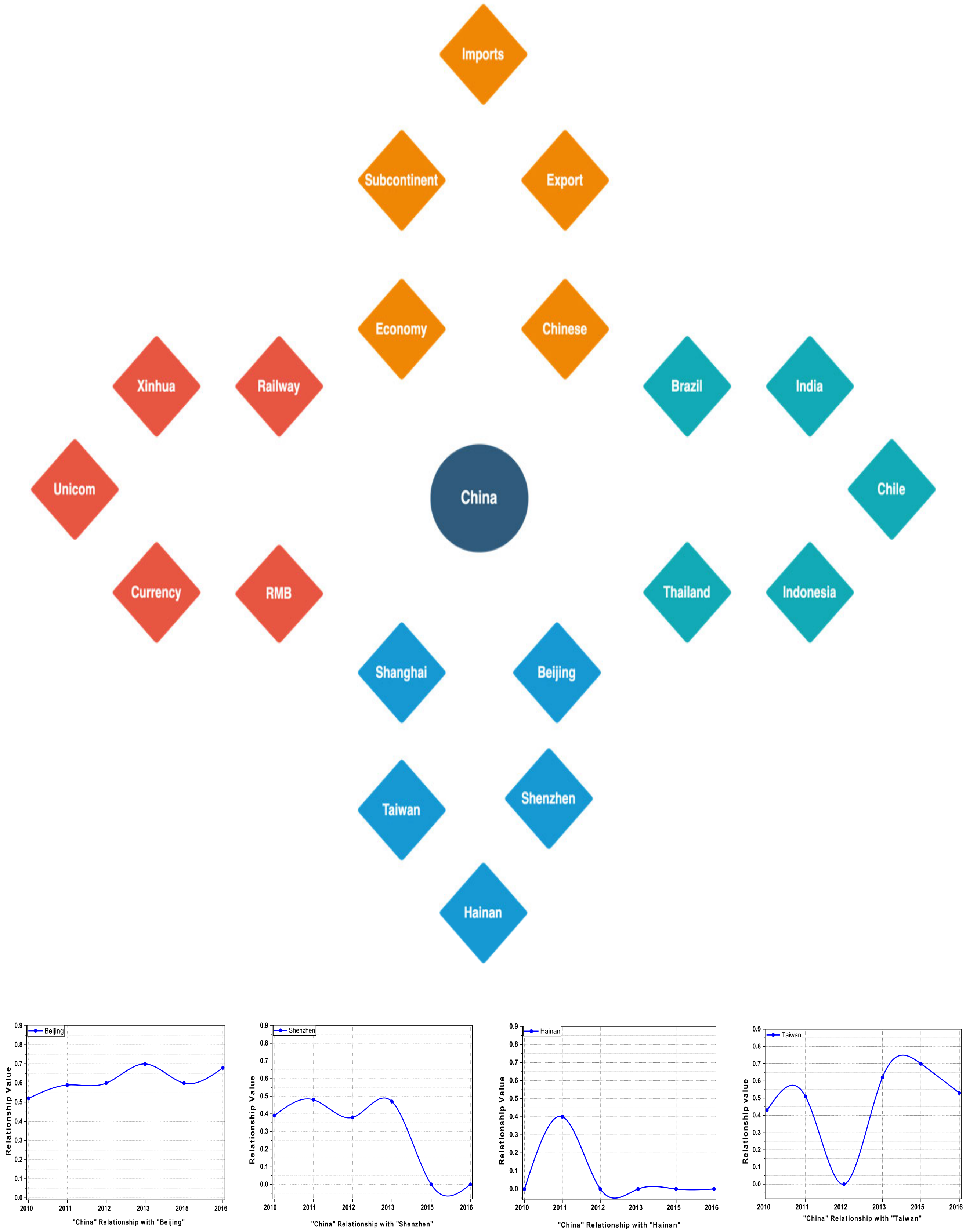

FIGURE 6. **"China" visual display model.**

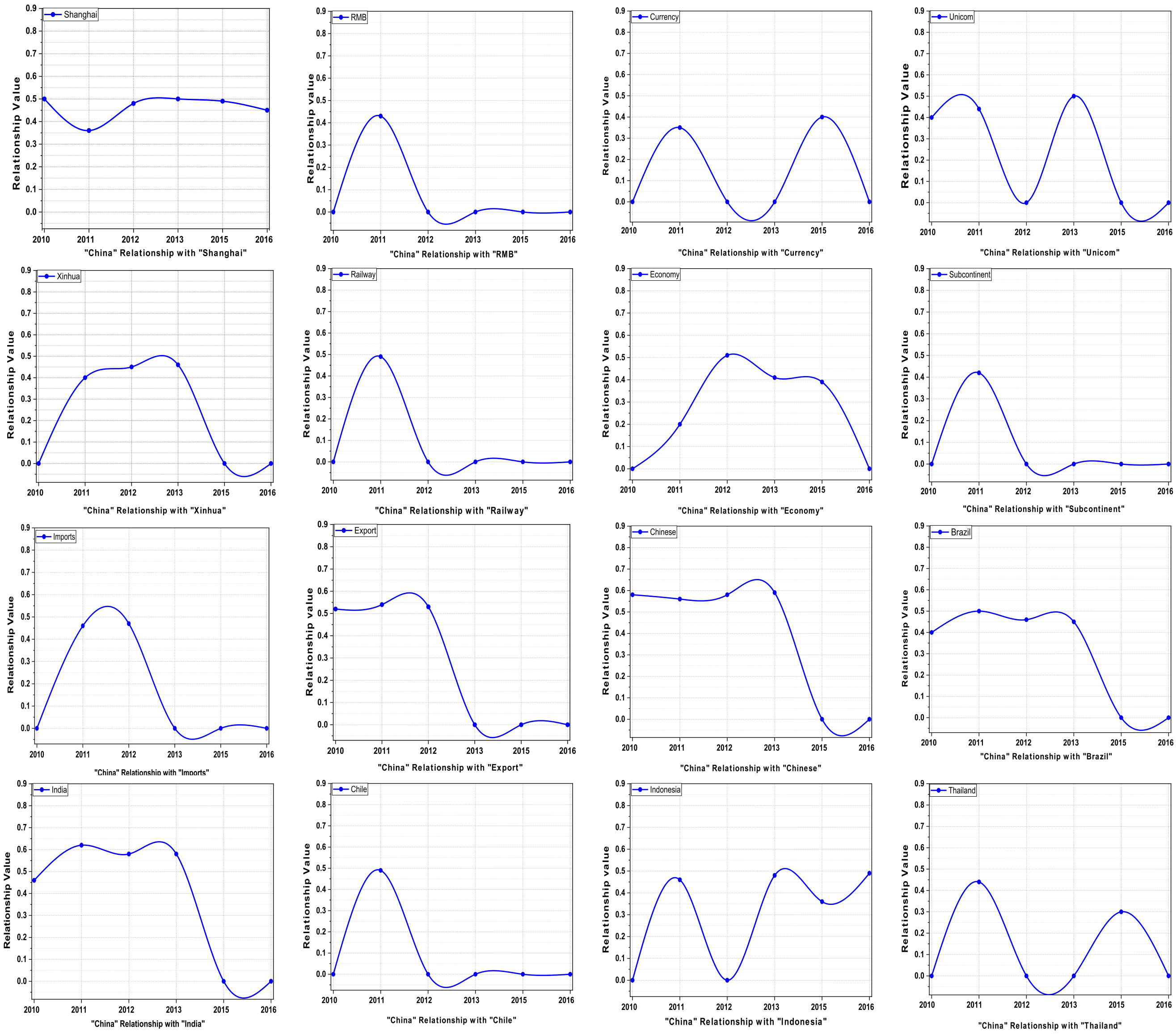

FIGURE 6. *(Continued.)* "China" visual display model.

fields. We have identified the top 20 trends related to "China" from different areas, including some countries, regions, and cities, as well as business terms such as "export," "agriculture," and "trade," based on the specific year when the data was collected. In addition, we have identified some general trending topics related to "China", such as "Xinhua" (a Chinese news agency) and "Unicom" (a mobile SIM company). By examining the visual model, one can quickly identify and understand the connections between these trends. This graphical representation enables users to gain insights into how these trends are interrelated and how they evolve over different time periods.

This experiment showcases the methodology's effectiveness in identifying trends not only in the technical field but also in various other domains. Initially, we identified the top 20 words associated with "China" in each year of the dataset. Subsequently, we constructed a visual model that facilitates comprehension for the general public. By clicking on each general trend related to "China", users can observe the graphical representation of its relationship percentage from 2010 to 2016 (Except 2014). This enables the identification of trends that experienced fluctuations, disappearances, and reappearances over time. Figure. 6 displays the names and graphical representations of the top 20 trends connected to "China" in each year of the dataset along with their corresponding relationship percentage fluctuations. The study provides a visual representation of the trends identified by our method across different fields related to "China". It demonstrates our ability to identify trends not specifically within a technological field but also in various general fields. Figure. 6 contains trends of "China" in diverse areas, including countries, regions, cities, and business terms such as "export," "agriculture," and "trade." These trends were identified based on the specific year when the data was collected.

To ensure better results and understanding, we have created a visualization model that allows a simple user to click and obtain results without requiring the expertise of a tech giant. This model offers an accessible and user-friendly way to explore trends related to "China", and it can help individuals and organizations stay up-to-date with the latest developments in the region. Moreover, this model can be applied to any general word or phrase in the future, providing flexibility to explore trends across different regions, fields, and periods. This can be an invaluable tool for businesses, governments, and researchers alike, providing them with real-time insights into the latest trends and developments in their field of interest.

The visualization model presented in this study offers an intuitive and easy-to-understand representation of the trends in text data streams. By providing a comprehensive view of the hot topics, the proposed visualization model enables users to quickly identify and track the most relevant trends. The interactive features of the model allow users to explore the data and gain deeper insights into the relationship between different topics. The proposed visualization model serves as a valuable tool for trend analysis in various fields, including business, social media, and news. In the next section, we will discuss our word2vec-based knowledge graph, which further enhances the effectiveness of our proposed method for trend detection in text data streams.

F. WORD2VEC-BASED KNOWLEDGE GRAPH

Continued growth of Internet technologies, and new-generation applications such as 5G, block-chain, and cloud computing are accelerating the process of digitalization and engineering advancement in several fields [39]. Such developments in the Internet have emerged problems like data sorting and obtaining information from data. Sequence or understanding of the information is as vital as information in the data. Knowledge representation and reasoning, inspired by human problem-solving represent knowledge for intelligent systems to gain the ability to solve complex tasks [40]. The concept of a knowledge graph was first defined as a large-scale knowledge base composed of a large number of entities and relationships between them proposed by "Google" in 2012 [41]. A research work describes a knowledge graph as a structure used to store and transmit real-world knowledge [42].

Word2Vec is a popular technique for natural language processing that can capture semantic relationships between words in large datasets. By using word2vec to generate embeddings, we can create a knowledge graph that maps out the relationships between different words and concepts. This can help us gain a deeper understanding of the underlying structure of the data and can be used to identify trends, discover patterns, and extract meaningful insights. In this section, we explore how we can leverage the power of word2vec to construct a knowledge graph from our news data and use it to gain new insights into the world of technology.

The experiment utilized data from businessinsider.com to create a knowledge graph on the given topic. However, the challenge lies in presenting the information in a way that is easily understandable to the general public, particularly when dealing with a large volume of streaming news data. Currently, there is a lack of effective methods and tools for tracking topics of interest and generating comprehensive knowledge graphs of information. To address this issue, we aim to provide a general approach that offers a visual knowledge graph of streaming data sources. This will benefit various users such as technical professionals, business experts, educators, students, researchers, international journals, conferences, and the general public, among others. By presenting information in the form of a visual knowledge graph, our proposed toolbox can help users obtain complete and accurate information about their target topic. Overall, this approach can help bridge the gap between technical jargon and general understanding, making it easier for users to access and digest information in a meaningful way.

The success of our experiment can be observed in Fig. 7, where we have visualized the knowledge graph of the targeted word "Apple" data. In our initial experiment, we considered Apple's top 5 trends in the 2010 news section, and the results were "iPad", "iPhone", "Nokia" "Sony", and "Siri". The visualization of the knowledge graph helps to illustrate the connections between these words. The purpose of knowledge graphs is to make it easier for the general public to understand while using a given tool. The knowledge graph above displays the relationship between Apple and the top 5 words related to it, consisting of actual relations (blue lines) and inferred relations (red and green lines). The word2Vec-based knowledge graph can serve as a powerful tool for discovering intricate connections between technology companies, tracking the evolution of market trends, and understanding the interplay between global events and their impact on specific industries. The practical implications of the knowledge graph are far-reaching, catering to various stakeholders. Business professionals can leverage it to make informed decisions about their strategies and investments. Researchers can utilize it for in-depth analysis and exploration of complex relationships within the data. Educators can incorporate it into their curricula to provide enriched learning experiences for students. Additionally, the general public can access user-friendly and comprehensive information about current topics, enabling them to stay well-informed and engaged with the latest developments.

As discussed, our method outperforms other methods in terms of its ability to generate results that are not limited to a specific field or industry. For instance, if a specific topic such as "LG and Apple" is given, the results are restricted to a particular field, as indicated in Figs. 4,5. However, if the topic

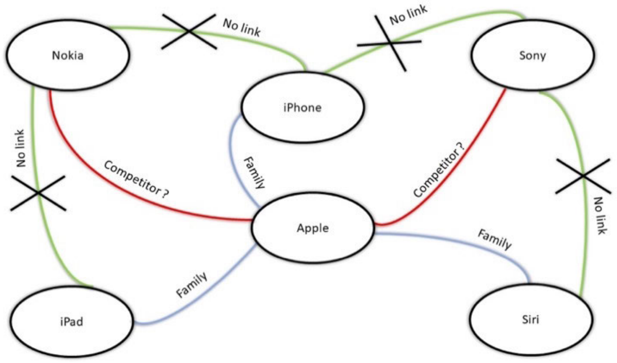

FIGURE 7. **Knowledge graph representation of "Apple" trends.**

is broad and general, such as "China", "America", "Internet", or any other general topic. The proposed framework can provide trends related to the topic across almost every aspect of daily life as graphed in Fig. 6. As a result, users and researchers can find almost all popular and relevant trends related to the topic and their relationships within seconds. Our proposed method which includes a novel model NSEM, visual display, and a semantic similarity-based knowledge graph can be successfully used for detecting hot topic trends, finding time series of the data segments, and providing a visual display of raw data.

The methodology and tools have shown promising results for analyzing streaming data sources and detecting trends. While our technical results are limited to the domain-specific language used in our dataset, the visualization model we developed for general words has broader implications. It can be applied to any text data stream, regardless of the field or industry, to detect trends and analyze the relationship between words. This versatility makes it a valuable tool for a wide range of applications, from business analytics to social media monitoring. Moreover, our visual models and knowledge graphs further enhance the interpretation of the semantic relationships between topic trends, making the information more accessible and engaging for users. These tools provide a comprehensive view of the evolving trends and correlations, enabling users to identify significant patterns and draw meaningful conclusions. By highlighting the practical applications of our proposed approach, we aim to underscore its potential value to various professionals in different domains. The method can aid in trend detection, topic analysis, and gaining valuable insights from large-scale news data. These contributions have significant implications for a range of applications in the field of natural language processing, including sentiment analysis, market research, and news recommendation systems, and provide meaningful context for business analysts, researchers, tech experts, and students.

G. COMPARISON TO STATE-OF-THE-ART APPROACHES

To provide a comprehensive understanding of the proposed method, it is essential to compare it with related research in the field. In recent years, various methods have been proposed for topic trend analysis, including those based on machine learning algorithms, data mining techniques, and natural language processing. However, there is a research gap regarding the utilization of distributed representations for trend detection, visual models for showcasing and understanding trends for the general public, word2vec-based knowledge graphs for topic trend identification, and the analysis of text data streams. Therefore, we have emphasized the importance of our work beyond the conventional state-of-the-art considerations and conducted a thorough and comprehensive comparison between our proposed method and previous research. Taking into account various aspects such as the data domain, approach, framework, chronological analysis, visualization model, and the addition of a word2vec-based knowledge graph, as shown in Table.1. By comparing and contrasting the proposed method with state-of-the-art approaches, it is possible to evaluate its effectiveness and highlight its contribution to the field.

Dynamic Topic Modeling (DTM) presented by Blei et al. [14], focuses on analyzing text data that evolves over time, explicitly modeling topic changes. In contrast, NSEM is designed for sequential data streams, such as streaming news, to identify hot topic trends and correlations as they evolve chronologically. DTM can handle large-scale corpora and provides fine-grained temporal analysis, while NSEM efficiently processes streaming data with flexibility in time granularities. While DTM is more versatile in various applications with temporal text analysis needs, NSEM caters specifically to real-time trend detection in streaming news data and similar sequential datasets. By leveraging distributed representation techniques like word2vec, NSEM can capture semantic relationships and patterns within the data, allowing us to identify and analyze trends that gain prominence or decline during diverse time periods. In summary, DTM and NSEM offer distinct approaches to understanding temporal patterns in textual data, with DTM providing a broader temporal view, and NSEM focusing on real-time trend detection and analysis in streaming datasets. Both DTM and NSEM diverge in their approaches, goals, and unique contributions to the field. Each model possesses distinct strengths and can be chosen based on specific research objectives and demands.

Rajapaksha et al. [29], Asgari-Chenaghlu et al. [26], Suryadi et al. [18], Behpour et al. [19], Lv et al. [20], Cavalli et al. [23], Ravi et al. [25], Azizi et al. [21], and Alkhodair et al. [37] have researched trend detection using different methods. However, none of their studies included the distributed representations method, stream text data, chronological analysis of the data, visualization models for topic trends, or distributed representations-based knowledge graphs. Our methodology presents significant advantages over previous approaches in the field of topic trend detection. By utilizing distributed representations like Word2Vec, we achieve a more nuanced understanding of word context, enhancing trend identification accuracy. Additionally, our approach includes chronological analysis, tracing the evolution of words over time to gain insights into topic changes. Unlike other studies, we handle streaming text data,

TABLE 1. Comparison of the proposed method with state-of-the-art approaches.

References	Data Domain	Approach	Framework	Chronological Analysis	Visualization Model	Word2Vec-based Knowledge Graph
Blei et al. [14]	Large Documents /Limited	Time Evolution	Specific	✓	×	×
Rajapaksha et al. [29]	Twitter/Limited	Detection	Specific	×	×	×
Asgari-Chenaghlu et al. [26]	Twitter/Limited	Detection	Specific	×	×	×
Suryadi et al. [18]	Customer Reviews / Limited	Detection	Specific	×	×	×
Behpour et al. [19]	Time Biased Clustering / Limited	Detection	Specific	×	×	×
Lv et al. [20]	Categorical Data / Limited	Detection	Specific	×	×	×
Cavalli et al. [23]	Multivariate Data / Limited	Detection	Specific	×	×	×
Ravi et al. [25]	Healthcare System / Limited	Detection	Specific	×	×	×
Azizi et al. [21]	Twitter / Limited	Detection	Specific	×	×	×
Alkhodair et al. [37]	Social Media / Limited	Detection	Specific	×	×	×
Proposed Method	**Stream Text News / Unlimited**	Detection	**General**	✓	✓	✓

enabling real-time trend analysis. Moreover, our visualization model provides an intuitive representation of topic trends, making complex findings easily understandable. Furthermore, we construct knowledge graphs based on distributed representations, offering a comprehensive view of word relationships. These innovative features collectively strengthen the validity and impact of our research, making it a valuable contribution to knowledge discovery and natural language processing in various domains.

Additionally, our method is versatile and capable of identifying trends not only in a specific field but also in diverse general domains. While these approaches focus on trends within a particular field, our methodology is designed to detect trends across different fields, including both technical and general areas. Furthermore, Suryadi et al. [18] and Azizi et al. [21] proposed frameworks limited to specific research areas and data domains and lacking future applicability beyond the COVID-19 pandemic. In contrast, the proposed study goes beyond previous research by utilizing chronological analysis, developing a word2vec-based visualization model, and creating a distributed representations-based knowledge graph. These advancements provide a more comprehensive and versatile framework for trend detection and analysis, with future applicability beyond the current pandemic. While previous research has contributed to the field of trend detection, the proposed study represents a significant advancement in terms of methodology and future applicability.

H. THREATS TO VALIDITY

Several potential threats to the validity of the proposed method should be considered. One potential threat to internal validity can be the use of word2vec models. Although word2vec is a widely used and well-established method for natural language processing, it is not without limitations. For example, word2vec models may struggle with rare or unseen words, leading to inaccuracies in the results. Additionally, the quality of the embeddings generated by the models may vary depending on the size and quality of the training data used. As the size of the data news from the year 2014 was small, the performance of the experiments conducted on it was not satisfactory. To mitigate this issue, we made a careful decision to remove the dataset of the year 2014 from our experiments. This step allowed us to avoid potential pitfalls associated with the small dataset size and improve the overall reliability of our findings. By focusing our analysis on the data from 2010 to 2016 (excluding 2014), we aimed to provide a more comprehensive and accurate assessment of the topic trends and semantic similarity relationships.

Another potential threat to internal validity is the use of NSEM. While NSEM is an effective method for analyzing sequential periods in data streams, it is still a relatively new technique, and its effectiveness may vary depending on the specific application and the characteristics of the data being analyzed. A potential threat to external validity is the generalizability of our results. To validate the effectiveness of the NSEM approach, we utilized the hold-out dataset technique. By randomly selecting a portion of our dataset as the "hold-out sample," which remained separate from the main data used for model training, we ensured an unbiased evaluation of the NSEM model's generalization capability. Applying the trained models to this unseen data allowed us to gauge their performance on new and diverse content, supporting the robustness of our proposed approach. While we have demonstrated the effectiveness of our approach on several datasets, it is possible that our results may not generalize to other datasets or domains.

Additionally, our results may be influenced by factors such as the specific preprocessing techniques used or the choice of hyperparameters. Finally, there is a potential threat to construct validity, which is the extent to which our measurements accurately reflect the underlying constructs of interest. While we have taken steps to ensure the validity of our measurements, other factors may influence the relationships between the variables that we have not accounted for. However, the proposed study has demonstrated the effectiveness of analyzing sequential periods in data streams and detecting hot topic trends. It is still important to consider the potential threats to validity when interpreting the results. By acknowledging and addressing these potential threats, we can ensure that our findings are robust and reliable.

I. DISCUSSION

In this research, we have presented the results of our analysis of sequential periods in data streams using distributed representations and NSEM. We have demonstrated that our approach effectively identifies and analyzes topic trends and hotspot relationships in different datasets. We have also shown that using NSEM improves the quality of our results, enabling us to identify the evolutionary process of correlations over time. Additionally, we have constructed a visual display model for general words, making our method generalizable and applicable to a wide range of categories. We have also discussed the potential applications of our model in fields such as finance where the ability to accurately predict market trends is crucial, and healthcare, where it could have significant implications for improving patient outcomes highlighting the importance of accurate trend prediction and relationship analysis in these areas. In addition to these fields, our model can be effectively applied in marketing and advertising to identify emerging consumer trends and adapt strategies accordingly. Social media analysis can benefit from our model to track trending topics and sentiments, while e-commerce platforms can use it for inventory management and product recommendations. News media can leverage our model to understand public interests and deliver relevant content, and education settings can utilize it to identify emerging research topics. In the corporate world, our model aids in market trend tracking and competitor analysis for informed decision-making. Government and policy analysis can benefit from monitoring public sentiments, while sports analytics can use our model to identify emerging sports-related topics and player performances. The flexibility and versatility of our model enable its application in various domains, where timely and accurate trend detection is crucial for gaining insights and making informed decisions.

However, we have also acknowledged the potential threats to the validity of our study, including the limitations of word2vec and the generalizability of our results. By acknowledging these limitations, we can ensure that our findings are interpreted and contextualized appropriately.

V. CONCLUSION

This research has demonstrated how to accurately detect and identify topic trends and analyze hotspot relationships intelligently using distributed representations. The proposed model displayed a linear structure that enables precise semantic relations. Secondly, we chronologically analyzed the differences in sequential periods of different datasets, making it capable of analyzing the relationships between words and taking into account how those relationships changed over different time periods. We successfully built several separate word2vec models for each dataset, instead of only one model. We also thank NSEM for its great improvement in the quality of our work, allowing us to find the evolutionary process of correlations in each model. NSEM resulted in faster training and significantly better representations of uncommon datasets. For the first time, we explain the insides of NSEM, which are crucial in analyzing the difference between sequential periods. Additionally, we constructed a visual display model for common words, making the method general and applicable to any category Another vital outcome of this work is constructing a knowledge graph of topic trends using semantic similarity through word2vec. The presened toolbox plays a pivotal role in providing valuable insights, facilitating data-driven decision-making, fostering research advancements, enhancing educational experiences, and empowering the general public with comprehensive and user-friendly information about current topics, making our research indispensable with multifaceted applications. Therefore, this work can be viewed as complementary to existing methods.

In future work, we shall investigate the application of our model to other types of data streams, such as social media or sensor data. More advanced visualization techniques can be used to help users better understand and interpret the results and different clustering algorithms can be investigated to improve the accuracy of our topic identification and relationship analysis. Potential applications of our model can be explored in fields such as finance, where the ability to accurately predict market trends is crucial, and in healthcare, it could have significant implications for improving patient outcomes.

DATA AVAILABILITY

The datasets related to the manuscript can be accessed at (https://github.com/ZohaibAhmadKhan/NSEM-Dataset) for public use currently. However, for additional materials or information, interested individuals can request access from the corresponding author and such requests will be considered reasonable.

REFERENCES

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.

[2] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2013, pp. 746–751.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[4] Z. A. Khan, Q. Wang, Y. Liu, and Y. Li, "Streaming news sequential evolution model based on distributed representations," in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 9647–9650.

[5] G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," *J. Supercomput.*, vol. 77, no. 11, pp. 12320–12335, Nov. 2021.

[6] R. Raja, P. C. Sharma, M. R. Mahmood, and D. K. Saini, "Analysis of anomaly detection in surveillance video: Recent trends and future vision," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 12635–12651, Mar. 2023.

[7] F. Gurcan, G. G. M. Dalveren, N. E. Cagiltay, and A. Soylu, "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," *Sci. Educ.*, vol. 32, pp. 1–24, Feb. 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," *Educ. Inf. Technol.*, vol. 28, no. 3, pp. 3557–3579, Mar. 2023.

[10] Y.-E. Kim, M.-G. Kim, and H. Kim, "Detecting IoT botnet in 5G core network using machine learning," *Comput., Mater. Continua*, vol. 72, no. 3, pp. 4467–4488, 2022.

[11] J. Bao, Y. Chen, J. Yin, X. Chen, and D. Zhu, "Exploring topics and trends in Chinese ATC incident reports using a domain-knowledge driven topic model," *J. Air Transp. Manage.*, vol. 108, May 2023, Art. no. 102374.

[12] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, "Bangla-BERT: Transformer-based efficient model for transfer learning and language understanding," *IEEE Access*, vol. 10, pp. 91855–91870, 2022.

[13] J. A. Khan, Y. Xie, L. Liu, and L. Wen, "Analysis of requirements-related arguments in user forums," in *Proc. IEEE 27th Int. Requirements Eng. Conf. (RE)*, Sep. 2019, pp. 63–74.

[14] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[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, Dec. 2022.

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

[17] F. Gurcan, G. G. M. Dalveren, N. E. Cagiltay, D. Roman, and A. Soylu, "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 COVID-19 pandemic," in *Proc. Int. Vis., Informat. Technol. Conf. (IVIT)*, Nov. 2022, pp. 172–178.

[19] S. Behpour, M. Mohammadi, M. V. Albert, Z. S. Alam, L. Wang, and T. Xiao, "Automatic trend detection: Time-biased document clustering," *Knowl.-Based Syst.*, vol. 220, May 2021, Art. no. 106907.

[20] F. Lv, T. Liang, J. Zhao, Z. Zhuo, J. Wu, and G. Yang, "Latent Gaussian process for anomaly detection in categorical data," *Knowl.-Based Syst.*, vol. 220, May 2021, Art. no. 106896.

[21] F. Azizi, H. Hajiabadi, H. Vahdat-Nejad, and M. H. Khosravi, "Detecting and analyzing topics of massive COVID-19 related tweets for various countries," *Comput. Electr. Eng.*, vol. 106, Mar. 2023, Art. no. 108561.

[22] M. Yukselen, A. Mutlu, and P. Karagoz, "Predicting the trending research topics by deep neural network based content analysis," *IEEE Access*, vol. 10, pp. 90887–90902, 2022.

[23] S. Cavalli and M. Amoretti, "CNN-based multivariate data analysis for Bitcoin trend prediction," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107065.

[24] Z. Guo, K. Yu, A. Jolfaei, A. K. Bashir, A. O. Almagrabi, and N. Kumar, "Fuzzy detection system for rumors through explainable adaptive learning," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3650–3664, Dec. 2021.

[25] V. Ravi, T. D. Pham, and M. Alazab, "Attention-based multidimensional deep learning approach for cross-architecture IoMT malware detection and classification in healthcare cyber-physical systems," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1597–1606, Aug. 2023.

[26] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, L. Farzinvash, M.-A. Balafar, and C. Motamed, "TopicBERT: A cognitive approach for topic detection from multimodal post stream using BERT and memory–graph," *Chaos, Solitons Fractals*, vol. 151, Oct. 2021, Art. no. 111274.

[27] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 sensing: Negative sentiment analysis on social media in China via BERT model," *IEEE Access*, vol. 8, pp. 138162–138169, 2020.

[28] P. Maddigan and T. Susnjak, "Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45181–45193, 2023.

[29] P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT, XLNet or RoBERTa: The best transfer learning model to detect clickbaits," *IEEE Access*, vol. 9, pp. 154704–154716, 2021.

[30] C.-O. Truică, E.-S. Apostol, and A. Paschke, "Awakened at CheckThat! 2022: Fake news detection using BiLSTM and sentence transformer," in *Proc. Working Notes CLEF*, 2022, pp. 1–9.

[31] K.-L. Chiu, A. Collins, and R. Alexander, "Detecting hate speech with GPT-3," 2021, *arXiv:2103.12407*.

[32] J. D. Nunes, M. Carvalho, D. Carneiro, and J. S. Cardoso, "Spiking neural networks: A survey," *IEEE Access*, vol. 10, pp. 60738–60764, 2022.

[33] R. Catelli, L. Bevilacqua, N. Mariniello, V. S. D. Carlo, M. Magaldi, H. Fujita, G. De Pietro, and M. Esposito, "A new Italian cultural heritage data set: Detecting fake reviews with BERT and ELECTRA leveraging the sentiment," *IEEE Access*, vol. 11, pp. 52214–52225, 2023.

[34] S. S. Sabry, T. Adewumi, N. Abid, G. Kovács, F. Liwicki, and M. Liwicki, "HaT5: Hate language identification using text-to-text transfer transformer," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.

[35] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proc. IEEE*, vol. 111, no. 5, pp. 465–499, May 2023.

[36] J. A. Khan, L. Liu, L. Wen, and R. Ali, "Crowd intelligence in requirements engineering: Current status and future directions," in *Requirements Engineering: Foundation for Software Quality*. Essen, Germany: Springer, Mar. 2019.

[37] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102018.

[38] X. Rong, "Word2vec parameter learning explained," 2014, *arXiv:1411.2738*.

[39] J. Li, H. Zhang, and Z. Wei, "The weighted Word2vec paragraph vectors for anomaly detection over HTTP traffic," *IEEE Access*, vol. 8, pp. 141787–141798, 2020.

[40] L. Juanzi and H. Lei, "Reviews on knowledge graph research," *J. Shanxi Univ., Natural Sci. Ed.*, vol. 40, no. 3, pp. 454–459, 2017.

[41] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[42] A. Hogan et al., "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, 2021.

ZOHAIB AHMAD KHAN received the master's degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree in control science and engineering. His research interests include machine learning, natural language processing, data mining, knowledge graphs, and artificial intelligence. He aspires to harness the power of text data for developing intelligent systems and advancing our understanding of these fields.

YUANQING XIA (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from Beihang University, Beijing, China, in 2001. He was a Research Fellow in several academic institutions, from 2002 to 2008, including the National University of Singapore and the University of Glamorgan, U.K. Since 2004, he has been with the Beijing Institute of Technology (BIT), China, where he is currently a Full Professor and the Dean of the School of Automation. He is also the Director of the Specialized Committee on Cloud Control and Decision, Chinese Institute of Command and Control (CICC). He has published 16 monographs in Springer, John Wiley, and CRC, and more than 500 papers in international scientific journals. He has been a highly cited scholar, since 2014, by Elsevier. His research interests include cloud control systems, networked control systems, robust control, signal processing, active disturbance rejection control, and flight control. He is a member of the 8th Disciplinary Review Group, Academic Degrees Committee, State Council, and the Big Data Expert Committee of the Chinese Computer Society. He was granted by the National Outstanding Youth Foundation of China, in 2012, and was honored as the Yangtze River Scholar Distinguished Professor, in 2016, and the leading talent of the Chinese ten thousand talents program. He received the Second Award of the Beijing Municipal Science and Technology (No. 1), in 2010 and 2015, the Second National Award for Science and Technology (No. 2), in 2011, the Second Natural Science Award of the Ministry of Education (No. 1), in 2012 and 2017, and the Second Wu Wenjun Artificial Intelligence Award, in 2018 (No. 1). More than five of his students have obtained excellent doctoral thesis awards from the Chinese Association of Automation or the Chinese Institute of Command and Control. He is the Vice-Chairperson of the Internet of Things Working Committee, Chinese Institute of Instrumentation. He is the Deputy Editor of the *Journal of Beijing Institute of Technology*. He is an Associate Editor of *Acta Automatica Sinica*, *International Journal of Automation and Computing*, *Guidance, Navigation, and Control: Theory and Applications*.

SHAHZAD ALI received the master's degree in control theory and control engineering from North China Electric Power University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree in control science and engineering with the Beijing Institute of Technology, Beijing. His research interests include machine learning, load frequency control, and control algorithms.

JAVED ALI KHAN received the Ph.D. degree in software engineering from Tsinghua University (QS ranked 13th), Beijing, China. He is currently a Senior Lecturer with the Foundation of Software Engineering (FSE) Group, Department of Computer Science, University of Hertfordshire, U.K. Previously, he was an Assistant Professor and the Cum Chairperson with the Department of Software Engineering, University of Science and Technology Bannu, Pakistan. He has published more than 40 papers in reputable journals and conferences. His research interests include requirements engineering, CrowdRE, argumentation and argument mining, mining software repositories, human values in software, quantum software engineering, feedback analysis, empirical software engineering, sentiment and opinion mining, requirements prioritization, mining fake reviews, sarcasm detection, and health analytics.

S. S. ASKAR received the B.Sc. degree in mathematics and the M.Sc. degree in applied mathematics from Mansoura University, Egypt, in 1998 and 2004, respectively, and the Ph.D. degree in operation research from Cranfield University, U.K., in 2011. In 2012, he joined King Saud University, where he is currently a Professor with the Department of Statistics and Operation Research. He has been an Associate Professor with Mansoura University, since 2016. His research interests include game theory and its applications that include mathematical economy, dynamical systems, and network analysis.

MOHAMED ABOUHAWWASH received the B.Sc. and M.Sc. degrees in statistics and computer science from Mansoura University, Mansoura, Egypt, in 2005 and 2011, respectively, the Ph.D. degree in statistics and computer science in a channel program from Michigan State University, East Lansing, MI, USA, in 2015, and the Ph.D. degree from Mansoura University in 2015. In 2018, he was a Visiting Scholar with the Department of Mathematics and Statistics, Faculty of Science, Thompson Rivers University, Kamloops, BC, Canada. He is currently with the Computational Mathematics, Science, and Engineering (CMSE), Biomedical Engineering (BME), Radiology, Institute for Quantitative Health Science and Engineering (IQ), Michigan State University. He is an Associate Professor with the Department of Mathematics, Faculty of Science, Mansoura University. His current research interests include evolutionary algorithms, machine learning, image reconstruction, and mathematical optimization. He was a recipient of the Best Master's and Ph.D. Thesis Awards from Mansoura University, in 2012 and 2018, respectively.

NORA EL-RASHIDY received the M.Sc. and Ph.D. degrees from the Faculty of Computer Science and Information, Mansoura University, Egypt, in 2016 and 2020, respectively. She is currently an Associate Professor with the Machine Learning and Information Retrieval Department, Faculty of Artificial Intelligence, Kafrelsheikh University. She is an Assistant Professor with Galala University and the Machine Learning and Information Retrieval Department, Kafrelsheikh University; and a Manager of the Measurement and Assessment Central Unit. She is also a Freelancer, a Net and Share Point Developer, and a ASP Net Developer to finally have ten years of successful professional work experience. Her research interests include machine learning, medical informatics, distributed and hybrid clinical decision support systems, big data, and cloud computing. She is very interested in disease diagnosis and treatment research. She is a reviewer for many journals (Research Gate: https://www.researchgate.net/profile/Nora-El-Rashidy; Scopus profile: https://www.scopus.com/authid/detail.uri?authorId =57211502814.)