Research on Weak Signal Recognition Based on LDA-BERT Fusion Model [Retracted due to Suspected Serious Academic Misconduct] Postprint

Authors: Yang Bo, Shao Wanting

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] To address the inadequacy of existing research on fully automatic weak signal recognition, this study proposes a fully automatic weak signal recognition method based on an LDA-BERT fusion model.

[Method/Process] The approach utilizes an unsupervised LDA topic model to conduct topic classification on text datasets, constructs a dual-layer filtering function for topics and terms to extract early warning signals from the topic classification results, evaluates topic weakness through three metric functions: closeness centrality, topic weight, and topic autocorrelation, and extracts weak signals based on the normalized frequency and probability of terms within topics. Finally, the BERT deep learning model is employed to expand the context of weak signals and their similar words at the semantic level.

[Results/Conclusion] Using the severe outbreak event in early January 2021 as a case study, the constructed system model is validated using social media news datasets from the three months prior to the outbreak. Experimental results indicate that the method can effectively detect relevant weak signals and reveal the evolutionary characteristic of weak signals gradually intensifying over time. Moreover, while achieving fully automatic weak signal recognition, this fusion model demonstrates stronger result interpretability compared to single models.

Full Text

Research on Weak Signal Recognition Based on an LDA-BERT Fusion Model

Yang Bo[1],[2], **Shao Wanting**[1],[2] [1] School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330013 [2] Institute of Infor-

mation Resources Management, Jiangxi University of Finance and Economics, Nanchang 330013

Abstract: [Purpose/Significance] To address the limitations in existing fully automated weak signal recognition research, this paper proposes a fully automated weak signal identification method based on an LDA-BERT fusion model. [Method/Process] Using the unsupervised LDA topic model, text datasets are classified by topic. A double-layer filtering function for topics and terms is constructed to extract early warning signals from the topic classification results. The "weakness" of topics is evaluated through three metric functions: closeness centrality, topic weight, and topic autocorrelation. Weak signals are then extracted based on the normalized frequency and probability of terms within topics. Finally, the BERT deep learning model is employed to expand weak signal contexts and similar words from a semantic perspective. [Results/Conclusions] Taking the COVID-19 outbreak in early January 2021 as an example, the constructed system model was validated using social media news datasets from the three months preceding the outbreak. Experimental results demonstrate that this method can effectively detect relevant weak signals and uncover their evolutionary characteristic of gradually strengthening over time. Moreover, this fusion model not only achieves fully automated weak signal identification but also exhibits stronger result interpretability compared to single models.

Keywords: weak signals; LDA-BERT fusion model; COVID-19 epidemic **Classification Number:** G250 **DOI:** 10.13266/j.issn.0252-3116.2021.16.011

---

1 Introduction

In the era of big data, decision-making increasingly relies on the analysis of acquired data and information rather than intuition and experience alone. Competitive intelligence, as the foundation for data and information analysis, is crucial for its acquisition, collection, and identification. Weak signals constitute an important component of forward-looking research in competitive intelligence, providing valuable references for decision-makers to predict future opportunities and risks. Like most information, weak signals are extracted from massive datasets and form valuable intelligence through reasonable inference and connection. However, due to their predictive nature, they are also called early warning signals. Ignoring weak signals means disregarding or even suppressing warning signs that could prevent erroneous decisions—akin to running a red light while driving, which inevitably leads to failure. Therefore, research on weak signal identification holds practical significance for enabling decision-makers to timely perceive market opportunities and threats and formulate management strategies conducive to long-term development.

Currently, identifying weak signals and predicting future scenarios has become a goal for many researchers. Numerous techniques have been used to gain maximum insight from words or documents, but most require assistance from human

experts. Traditional topic modeling techniques have demonstrated their fully automated capabilities. Consequently, this study employs a well-known topic model: Latent Dirichlet Allocation (LDA). LDA is an unsupervised machine learning technique that can operate independently based on input document collections and specified topic numbers without requiring manually labeled training sets, enabling fully automated extraction of topics and corresponding keywords from datasets during weak signal identification.

However, LDA topic model extraction results are not all weak signals; they also include strong signals with clear classification orientation and noise signals that cannot reveal specific meanings. Therefore, further filtering is needed to detect hidden, important words identified as weak signals. Meanwhile, due to the rare characteristics of weak signals, the number of extracted weak signals is limited. To enable sufficient automated detection of weak signals, this study proposes a weak signal fully automated identification method based on an LDA-BERT fusion model. This approach constructs topic and term double-layer filtering functions to extract early warning signals from LDA topic classification results, evaluates topic weakness through three metric functions (closeness centrality, topic weight, and topic autocorrelation), and extracts weak signals based on normalized frequency and probability of terms within topics. Finally, to compensate for LDA' s bag-of-words limitations and enhance model result interpretability, the BERT method is applied to predict contexts for each filtered topic document, obtaining more words semantically related to weak signals.

2 Related Research

The concept of "weak signals" was first proposed by H. Igor Ansoff in 1975, defined as "symptoms of future possible changes." He considered weak signals as incomplete warnings from external or internal sources whose impacts cannot be accurately estimated. Organizations must prepare in advance to respond to signs of potential threats and opportunities in uncertain environments. Subsequently, scholars such as B. Coffman, P. Rossel, and S. Mendona supplemented the concept, identifying weak signals as: difficult to track and distinguish from noise; trivial and easily overlooked yet potentially impactful; and early clues to future changes and trends.

Chinese research on weak signals started later but has offered profound insights. Shen Guchao argued that weak signals enable early judgment of future trends through observation of signs in an organization' s competitive environment and analysis of industry personnel opinions. Shan Bin summarized four reasons for the "weakness" of weak signals: (1) few perceivable weak signals; (2) difficulty capturing effective information; (3) coexistence of misleading or false signals with effective information; and (4) limited cost and energy for signal collection. Zhao Xiaokang noted four main characteristics: gradual prominence during development; increasing certainty; progressively enriched effective information;

and continuously enhanced intelligence value as decision-making basis.

Currently, weak signal identification lacks automation, with most research relying on manual input or expert opinion. For instance, I. Griol-Barres et al. used heterogeneous and unstructured information from scientific, news, and social sources for quantitative weak signal detection, applying multi-word co-occurrence analysis to manually selected keywords and extracting accurate results through natural language processing. J. Yoon proposed a text mining-based weak signal topic identification method under the premise of expert-provided keywords, demonstrating feasibility through web news reports on solar cells. Deng Shengli et al. quantitatively identified weak signals through Analytic Hierarchy Process and membership functions under expert-given coefficients. These methods require substantial manual effort and yield results with strong subjective characteristics.

Meanwhile, scholars have endeavored to apply techniques such as deep learning and neural networks for predictive analysis of increasingly abundant online text data. Natural Language Processing (NLP) can effectively extract insights from text data, with word embedding technologies accurately capturing word similarity and context-based prediction. B. DienAdj et al. proposed an embedded topic model combining conventional topic models with word embeddings, though these techniques provide better results with labeled data compared to unlabeled data. In weak signal detection from web articles, text data typically lacks labels, making deep learning-based NLP techniques unable to ensure complete automation of the weak signal detection process.

Fully automated weak signal detection research remains in its infancy, with limited papers and projects. Topic models are widely applied for hidden information detection in fully automated identification processes. For example, L. Pépin used dynamic LDA to detect weak signals by applying LDA algorithms to texts from different time periods and using visualized scatter plots of topic evolution to detect weak signals. T. Gutsche proposed a method for automatic weak signal detection and forecasting using dynamic topic modeling and time series analysis, achieving good results. This study follows the same fully automated approach, selecting LDA to extract topics and corresponding keyword information from social media news datasets. Since LDA results include strong signals and noise signals besides weak signals, further filtering of LDA extraction results is necessary. Additionally, Zhuang Muni et al. noted LDA's bag-of-words limitation: in LDA, a document is merely a collection of words without sequential order, making it difficult to incorporate contextual information effectively. To address this, J. Maitre et al. proposed enhancing LDA with Word2Vec to obtain more weak signal-like words. However, L. Kahyun et al. found that BERT better captures semantic and grammatical complexity and is more helpful for polysemy resolution when comparing Word2Vec and BERT algorithms in NLP, making it superior for LDA model enhancement.

In summary, current weak signal identification methods have limitations, primarily relying on human experts during extraction and identification, lacking

research on fully automated weak signal identification, yielding limited extraction results, and having low interpretability and suboptimal early warning effects. Therefore, to achieve fully automated weak signal detection, compensate for LDA' s bag-of-words limitations, and enhance model result interpretability, this study introduces the BERT model for further processing and analysis of LDA extraction results, constructing an LDA-BERT fusion model. This model performs double-layer deep filtering on topics and terms while semantically expanding extracted weak signals to achieve better weak signal identification results.

---

3 Weak Signal Automatic Identification Method Framework

3.1 Method Overview

To reduce human expert intervention, this study designs a fully automated weak signal identification method using unsupervised text mining techniques related to topic modeling. LDA is commonly used to extract trending topics from text datasets. Unlike keyword-based weak signal detection research, topic models consider word meanings rather than words themselves. This paper employs the LDA topic model to find topics that may lead to weak signals but does not accept that all topics contain weak signals or that all terms within topics are weak signals. Besides weak signals, strong signals with clear classification orientation and noise signals without specific meaning still exist. Therefore, this paper proposes a double-layer filtering model for topic and term filtering to extract only potential weak signals.

The topic filtering model constructs a topic weakness evaluation function based on closeness centrality, topic weight, and topic autocorrelation to extract topics potentially containing weak signals. The term filtering model further extracts weak signal-related terms from these topics, primarily relying on normalized frequency and probability of terms within topics. However, due to the rare characteristics of weak signals, the number of extracted weak signals is small, making it difficult to discover their associations and resulting in low model interpretability. To solve this problem, referencing J. Maitre et al.' s method of enhancing extraction results with Word2Vec, this study adopts the superior BERT deep learning model to semantically expand weak signals comprehensively from context, obtaining more interpretable early warning information.

The method framework is as follows: (1) Data collection: This study collects social media news content over a period as input for weak signal identification research. (2) Weak signal identification: Includes data preprocessing and weak signal filtering. Data preprocessing involves stop-word removal and tokenization of collected text sets. Weak signal filtering includes using the LDA topic model to identify topics and filtering extracted topics and terms to find poten-

tial weak topics and weak signals. (3) Weak signal output: The BERT model word embedding is used to enhance identified weak signals and output them. The process is shown in Figure 1 [Figure 1: see original paper].

3.2 Data Collection and Preprocessing

In weak signal identification tasks, text dataset quality directly affects weak signal detection accuracy and foresight. This study uses Python tools for data collection and preprocessing, with basic steps as follows:

(1) Text data collection: Using web crawler technology to collect news data from the internet over a period. This study focuses on social media news due to its wide dissemination, strong timeliness, and fast propagation speed, making it optimal data for weak signal identification.

(2) Text set cleaning and tokenization: The collected news dataset is cleaned based on a Chinese stop-word list to filter out irrelevant, meaningless, and non-text information. The jieba tool is used for tokenizing cleaned data, producing a dataset ready for system input.

3.3 Weak Signal Automatic Identification Based on LDA-BERT Fusion Model

3.3.1 LDA Topic Model Training The LDA topic model, also known as Latent Dirichlet Allocation, finds latent and hidden information from text collections by maximizing word co-occurrence probability under a predetermined number of topics. For example, words like "football" and "sport" always co-occur and can be categorized under sports. D.M. Blei et al. demonstrated that LDA effectively extracts document topics. The main challenge of LDA is determining the optimal number of topics k: too many topics lead to insufficiently concentrated distributions with high inter-topic similarity, while too few topics result in overly broad content without clear classification orientation. Hyperparameters α and β represent document-topic density and word-topic density, respectively, playing important roles in establishing consistency between topics and terms.

Current mainstream methods for determining optimal topic number k include perplexity and coherence methods. Smaller perplexity values indicate better topic classification results, but Zhao Kai et al. found that perplexity values gradually decrease with increasing topic numbers, making it difficult to confirm the optimal k. Meanwhile, Huang Jiajia et al. proposed using coherence methods to balance topic quality, finding that topics extracted thereby have higher interpretability. This study follows this approach, applying the topic coherence measure c_v proposed by [**?**] to determine the optimal topic number.

To find the model with highest coherence, this study uses the control variable method, changing only the topic number k value per run while keeping other parameters constant. The c_v value is used as the coherence measure, determined

based on sliding window, normalized pointwise mutual information (NPMI), and cosine similarity, returning the topic number k with highest coherence as the optimal model result.

3.3.2 Topic Filtering The topic filtering function proposed in this section helps evaluate the likelihood of topics containing weak signals and filters topics extracted by the LDA topic model. This method is derived from the Logistic function, commonly used to illustrate population progress and growth but employed in linguistics to model language change. A marginal term' s propagation speed increases over time, but if it is a weak signal, it remains marginal after propagation speed increases. Based on this, this study creates a topic weakness evaluation function from two aspects: topic distribution characteristics and topic development characteristics, mining inherently weak topics that show weak association with other topics, low proportion in topic distribution, and long-term marginal status in development.

Three metric functions are defined to determine topic weakness: closeness centrality, topic weight, and topic autocorrelation function.

(1) Closeness centrality represents similarity through inter-topic distances. Many distance metrics can calculate similarity, such as Jaccard distance, cosine distance, and Hellinger distance. L. Pépin et al. found that distance measurements showing S-shaped changes most effectively represent text similarity. Based on this principle, this paper uses Hellinger distance to calculate topic z' s closeness centrality CC(z), where h represents Hellinger distance:

$$CC(z) = \sum_i h(z, z_i)$$

(2) Topic weight: Consistency among relevant topics within a model represents topic meaning assignment. Therefore, this paper defines topic z' s weight value W based on topic z' s consistency and the sum of all topics' consistency, where Coh(z) represents topic z' s consistency magnitude:

$$W(z) = \frac{Coh(z)}{\sum Coh(z_i)}$$

(3) Autocorrelation: Autocorrelation is a prevalent data trend analysis tool. Trend analysis predicts future possibilities based on historical data, quantifying and explaining trends and patterns in chaotic data over time. Autocorrelation describes relationships between the same variable across different periods, i.e., linear correlation between variable values and their lagged values. In news datasets, document frequencies related to certain topics change over time, so each topic' s autocorrelation across days helps filter topics potentially not containing weak signals. The autocorrelation function AC is defined as follows, where Cov(z)_k is topic z' s covariance

at lag k, and Var(z) is topic z' s variance:

$$AC(z) = \frac{Cov(z)_k}{Var(z)}$$

Using these three metric functions, the topic weakness evaluation function WK is constructed. Lower function values indicate weaker terms within topics, but when sufficiently low, they may be defined as noise. The weak function for topic z is defined as:

$$WK(z) = \frac{W(z) \times CC(z)}{1 + \exp(-AC(z))}$$

According to weak signal definitions, rarity is the main characteristic, and their movement is slow over time. Therefore, only topics corresponding to low WK function values are identified as weak topics. Based on the Pareto principle, weak signal information does not exceed 20%, and human experts define the noise threshold range as 0% to 2%, representing the probability of meaningless words in text. This paper decides to ignore low WK function values and defines new screening thresholds: noise below 1%, weak signals below 15%. Text signal distribution is shown in Figure 2 [Figure 2: see original paper].

3.3.3 Term Filtering While the defined topic filtering function can extract topics potentially containing weak signals, not all terms within these topics are weak signals. This section explores how to effectively extract weak signals from these terms. J. Chuang proposed uniqueness and saliency as term measurement criteria to judge information conveyed by terms in certain topics for understandable topics. Research found that saliency, generated by the difference between a word' s probability of being generated by latent topics and the topic' s marginal probability, is the product of overall frequency and uniqueness. Meanwhile, C. Sievert et al. sought the most relevant terms within topics through inter-term correlation, achieving better results than probability indicators.

Based on term probability within topics and correlation between terms and topics, this study proposes a new term filtering function PW(w), where NF(w) is term w's normalized frequency in topic z, and ϕ(w) represents term probability in topic w:

$$PW(w) = \frac{NF(w)}{1 + \exp(-(\phi(w) \times \log(\phi(w))))}$$

As described in Section 3.3.2, weak signals are rare; thus, this paper only extracts items with PW function values between 1% and 15%.

3.3.4 Weak Signal Output Under the two-layer filtering of topic and term filtering functions, weak signals can be well identified and extracted. However, result analysis and understanding are also crucial. Due to weak signals' rare characteristics leading to limited extraction, to further obtain words related to extracted weak signals and improve model interpretability, this study

uses the BERT deep learning model to enhance weak signal extraction results. BERT (Bidirectional Encoder Representations from Transformers), introduced by Google in 2018, uses the Transformer algorithm as its main framework to better capture bidirectional relationships in sentences. Through multi-task training objectives of Masked Language Model (MLM) and Next Sentence Prediction (NSP), the model achieves new heights in performance.

The BERT model structure is shown in Figure 3 [Figure 3: see original paper]. In the figure, Trm represents a Transformer encoder unit, CLS is a global vector indicating text start, SEP is a text separator, and MASK is a masked word. Each extracted weak signal word is treated as a vector to reconstruct word context, making words sharing common contexts in the corpus semantically close in space and expanding weak signals similar to extraction results. After inputting weak signal d_i{w_{ij}|j {1,2,⋯,m}} into the BERT model, each character in the weak signal is first vectorized into w_{ij}(ω+δ+) across three dimensions: word vector (ω), text vector (δ), and position vector ( ), then passed into the bidirectional Transformer encoder, finally outputting the vector set d' $i = \{w'$ {ij}|j {1,2,⋯,m}} that fuses full text semantic information.

This method resembles Word2Vec' s Skip-gram model, which predicts context information based on current words, but differs in that BERT considers more semantic and grammatical aspects than Word2Vec, with richer and more complete corpora, making it superior for word semantic expansion. Following previous scholars' research, this paper uses Google pre-trained set Fine-tuning, inputting each filtered weak signal into the BERT model. After training, it outputs a highly similar word list to the extracted weak signals, highlighting associations between weak signals extracted from news datasets and enhanced weak signals, obtaining stronger model interpretability.

---

4 Empirical Research

Weak signals occupy an important position in competitive intelligence, and many enterprises consider weak signal identification as a key development goal. This study applies the proposed LDA-BERT fusion model-based weak signal automatic identification method to social media news from Weibo and other platforms to detect early warning information for the COVID-19 outbreak in early January 2021. Using web crawler tools, 14,486 social media news articles were collected from November 1, 2020, to January 10, 2021. Python open-source libraries including jieba, Gensim, and others were used for tokenization, topic modeling, and natural language processing.

4.1 LDA Topic Model Training Results Analysis

To find the optimal topic number k for the best topic model, this study used Gensim' s LdaModel module and pyLDAvis visualization tool, comprehensively

evaluating different topic numbers through coherence measure c_v values and topic distribution conditions.

First, LDA topic modeling was performed on the preprocessed social media news dataset from November 1, 2020, to January 10, 2021. Second, the control variable method was used to measure coherence measure c_v values under different topic numbers k, with k ranging from 1 to 50. Finally, the optimal topic number k for the LDA topic model under the social media news dataset was selected by comprehensively analyzing coherence measure c_v values and topic distribution conditions under different k values. Model results are shown in Figure 4 [Figure 4: see original paper].

Higher topic model consistency indices indicate better classification results. As shown in Figure 4, when topic number k is 5 or 9, the model achieves higher consistency values. By comparing topic distribution conditions under different k values, when consistency indices are low (e.g., k = 20, 34, 50), topic distribution appears uneven with large inter-topic size differences. Therefore, through comprehensive analysis of consistency measure c_v values and topic distribution conditions, this paper determines the optimal topic number k for the social media news dataset to be 9.

4.2 Topic Filtering Results Analysis

For the nine topics extracted by the LDA topic model, the weakness of topics is evaluated through three metric functions—closeness centrality, topic weight, and topic autocorrelation—to filter out topics potentially containing weak signals.

This section first calculates Hellinger distances between each topic and all other topics, obtaining a 9$×$9 distance matrix to measure topic closeness centrality. Second, Gensim library is used to measure each topic' s daily document frequency, with partial data shown in Table 1 . Finally, topic autocorrelation functions are calculated based on all topics' daily document frequencies, where determining the lag period is critical. Generally, non-overlapping time series have lower autocorrelation than overlapping sequences, and less overlapping data yields lower autocorrelation. Since most trend analysis samples do not overlap, observing longer lag periods is beneficial.

Figures 5 [Figure 5: see original paper], 6 [Figure 6: see original paper], and 7 [Figure 7: see original paper] show topic filtering results for November 2020, December 2020, and January 2021, respectively. Shaded areas indicate topic filtering results potentially containing weak signals, with WK function values above 1% but below 15% of the result set.

In weak signal detection, this study aims to minimize topic filtering function values, i.e., maximize the denominator of the WK function. Therefore, a higher lag period is set to reduce overlapping periods between time series, minimizing the autocorrelation function AC. Thus, half of the observed data period is selected as the optimal time lag for the autocorrelation function, setting the lag

period to 15 days.

Using monthly observation cycles, three topics (T3, T7, T9) potentially containing weak signals are extracted from nine topics each month through the topic filtering function. However, not all terms within these topics are weak signals, so term filtering functions will further extract weak signals from them.

4.3 Term Filtering Results Analysis

The LDA topic model groups and ranks terms based on their occurrence probability within each topic. To capture weak signals within topics as much as possible, sufficient terms must be obtained from topics. Therefore, based on topic filtering results, 500 terms are extracted from topic T7 (November 2020), topic T3 (December 2020), and topic T9 (January 2021), with each term' s document frequency calculated. Term filtering functions are then applied to extract weak signals from them. Tables 2 , 3 , and 4 list weak signal extraction results for topics T7, T3, and T9, respectively.

Some words in the tables already show relevance to the outbreak. To enhance model interpretability, the BERT algorithm is applied to predict contexts for filtered terms, maximizing target word probabilities.

4.4 Weak Signal Extraction Results Analysis

This study aims to mine weak signals related to the COVID-19 outbreak in early January 2021. Therefore, social media news from the three months before the outbreak is used as the weak signal extraction dataset to gain maximum insight from social media news collections. To compensate for LDA' s bag-of-words limitations, obtain more words related to extracted weak signals, and enhance model result accuracy and interpretability, this study uses BERT deep learning to semantically expand filtered terms from context, endowing weak signals with more contextual information and similar words.

This study uses successful mining of early warnings for the outbreak as validity testing for the weak signal identification model. The word "continuous" in Table 2 is slightly relevant to the research content. After expansion, important weak signals such as "intensify," "rebound," "deteriorate," and "disease" were discovered. The specific expansion list is shown in Table 5 .

Combined with the actual outbreak background, starting January 10, 2021, asymptomatic infections increased sharply in Hebei and gradually spread to surrounding cities, plunging China' s epidemic situation back into crisis. Among extracted weak signals, early warnings related to epidemic development were discovered. For example, in December 2020 topic T3 filtered terms, weak signals like "continuous," "deteriorate," and "increase" began shifting toward words like "outbreak," "infection," and "resurgence." By early January 2021, topic T9 filtered weak signals had become "urgent," "severe," "threat," etc. The "Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia" clearly states that

the virus fears high temperatures—i.e., it is heat-sensitive but cold-resistant. Lower temperatures make epidemic situations difficult to control, a key factor in outbreak resurgence. Weak signal extraction results including "temperature" and "frozen" also foreshadowed the outbreak from temperature drop and virus transmission perspectives.

Furthermore, weak signal evolution over time shows two trends:

(1) **Strengthening of early weak signals**: As time progresses, some weak signal words show enhanced predictive degree. For example, "vaccine" extracted from T7 in November 2020 and "infection" from T9 in January 2021 establish relevance to epidemic and infectious diseases with transmission characteristics. Combined with weak signals indicating development degree like "outbreak," "deteriorate," and "resurgence," they demonstrate early warnings of deteriorating epidemic situations. These warnings strengthen over time, with signals like "severe" and "threat" in January 2021 T9 directly indicating severity. Decision-makers should comprehensively consider development trends over time. Upon discovering negative weak signals related to outbreak resurgence, timely response, protection, and control measures should be taken to ensure public safety. For positive weak signals like "potential" and "upgrade" from November 2020 T7, "capital" and "opportunity" from December 2020 T3, and "rise" and "innovation" from January 2021 T9, organizations should leverage them to explore future markets and grasp favorable policies.

(2) **Similar words to weak signals**: As mentioned, weak signals' rare characteristics lead to limited extraction, increasing difficulty in analyzing their association with future possibilities. Using BERT deep learning can greatly enrich detected weak signals. For example, "increase" and "strict prevention" from November 2020 T7 both expand to epidemic-related weak signals like "disease" and "epidemic," warning of early signals for outbreak resurgence. Such weak signals are further derivations of originally extracted weak signals, no longer limited by the original dataset but diverging weak signal associations through deep learning algorithms, achieving the transition from abstract weak signal description to concrete matters.

In summary, the LDA-BERT fusion model-based weak signal identification method proposed in this study effectively detects weak signals in social media news datasets from November 2020 to January 2021. Comprehensive analysis and understanding reveal that some weak signals gradually strengthen semantically over time, providing beneficial references for predicting the outbreak situation in early January 2021.

This model solves problems of high manual participation and strong subjectivity in current weak signal identification research, achieving fully automated weak signal detection and greatly reducing human expert time and costs. The LDA-BERT fusion model and double-layer filtering function are proposed to

reasonably expand weak signals semantically while ensuring only relevant weak signals are extracted, yielding highly interpretable results and providing new methods and ideas for weak signal detection in intelligence collection.

The method has the following advantages: (1) **Generalization**: Extracted weak signals are not specific to particular domains or topics but are warning information deserving attention within specified timeframes, allowing decision-makers to select relevant weak signals based on their needs. (2) **Automation**: No human intervention or keyword assistance is required during weak signal extraction; the method automatically detects weak signals from text. (3) **Scientific rigor**: The innovative double-layer filtering function filters topic classification results, avoiding manual screening subjectivity and making the process more scientific and standardized.

However, this study still has some limitations: (1) Since both weak signals and noise share characteristics of being ambiguous in meaning and slow-moving, text denoising is not fully accomplished. (2) While this study effectively filters some text noise by setting longer lag periods using autocorrelation, it may also filter out some valuable weak signals, preventing completely lossless extraction from text collections. Future research will focus on text denoising in weak signal identification to provide more accurate early warning information for decision-makers.

References

[1] Wu Jinhong, Zhang Fei, Ju Xiufang. Big data: Opportunities, challenges and countermeasures for enterprise competitive intelligence[J]. Journal of Intelligence, 2013, 32(1): 5-9.

[2] Shao Bo, Song Jiwei. Risk identification and ranking in anti-competitive intelligence early warning[J]. Information Studies: Theory & Application, 2007, 30(5): 642-645.

[3] WISSEMAH. Driving through red lights[J]. Long range planning, 2002, 35(5): 521-539.

[4] MUHLROTH C, GROTTLKE M. A systematic literature review of mining weak signals and trends for corporate foresight[J]. Journal of business economics, 2018, 88(5): 643-687.

[5] Jiang Tian, Liu Xiaoping, Liu Huizhou. Research on noise topic filtering method based on keyword relevance indicator (KRI) for LDA[J]. Library and Information Service, 2020, 64(3): 92-101.

[6] YOON J. Detecting weak signals for long-term business opportunities using text mining of Web news[J]. Expert systems with applications, 2012, 39(16): 12543-12550.

[7] COFFMAN B. Weak signal research, part I: introduction[EB/OL]. [2021-07-10]. http://legacy.mgtaylor.com/mgtaylor/jotm/winter97/jotmwinter97.htm.

[8] ROSSEL P. Weak signals as a flexible framing space for enhanced management and decision-making[J]. Technology analysis and strategic management, 2009, 21(3): 307-320.

[9] MENDONA S, PINA E C M, KAIVO-OJA J, et al. Wildcards, weak signals and organisational improvisation[J]. Futures, 2012, 44(3): 218-228.

[10] SANDRO M, GUSTAVO C, JOAO C. The strategic strength of weak signal analysis[J]. Futures, 2004, 36(2): 201-218.

[11] IGOR ANSOFF H. Managing strategic surprise by response to weak signals[J]. California management review, 1975, 18(2): 21-33.

[12] HOLOPAINEN M, TOIVONEN M. Weak signals: Ansoff today[J]. Futures, 2012, 44(3): 198-205.

[13] Shen Guchao. Signal analysis: Another important topic in competitive intelligence research[J]. Library and Information Service, 2009, 53(20): 11-15.

[14] Shan Bin. Analysis and empirical research of weak signals from a cognitive perspective[D]. Beijing: Academy of Military Medical Sciences, 2014.

[15] Zhao Xiaokang. Weak signals: Identification, detection and response[J]. Journal of Intelligence, 2010, 29(1): 159-163.

[16] GRIOL-BARRES I, MILLA S, CEBRIÁN A, et al. Detecting weak signals of the future: A system implementation based on text mining and natural language processing[J]. Sustainability, 2020, 12(19): 1-22.

[17] GRIOL-BARRES I, MILLA S, MILLE J. System implementation for detection of future weak signals using text mining[J]. Revista española de documentación científica, 2019, 42(2): e234.

[18] Deng Shenglin, Lin Yanqing, Wang Ye. Feature extraction and quantitative identification of enterprise competitive weak signals[J]. Library and Information Service, 2016, 60(10): 67-75.

[19] HIRSCHBERG J, MANNING C D. Advances in natural language processing[J]. Science, 2015, 349(6245): 261-266.

[20] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing[J]. Journal of engineering, 2018, 13(3): 55-75.

[21] DIENG A B, RUIZ F J R, BLEI D M. Topic modeling in embedding spaces[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 439-453.

[22] PÉPIN L, KUNTZ P, BLANCHARD J, et al. Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted Tweets[J]. Computers & industrial engineering, 2017, 112(2): 450-458.

[23] GUTSCHE T. Automatic weak signal detection and forecasting[D]. Enschede: University of Twente, 2018.

[24] Zhuang Muni, Li Yong, Tan Xu, et al. Simulation of COVID-19 epidemic network public opinion evolution based on BERT-LDA model[J]. Journal of System Simulation, 2021, 33(1): 24-36.

[25] MAITRE J, MÉNARD M, CHIRON G, et al. A meaningful information extraction system for interactive analysis of documents[C]//2019 international conference on document analysis and recognition. Sydney: IEEE. 2019. 92-99.

[26] LEE K, FILANNINO M, UZUNER Ö. An empirical test of GRUs and deep contextualized word representations on de-identification[J]. Studies in health technology and informatics, 2019, 264: 218-222.

[27] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003(3): 993-1022.

[28] Zhao Kai, Wang Hongyuan. Research on LDA optimal topic number selection: Taking CNKI literature as an example[J]. Statistics & Decision, 2020, 36(16): 175-179.

[29] CHANG J, GERRISH S, WANG C, et al. Reading tea leaves: How humans interpret topic models[C]//Neural information processing systems. New York: Curran Associates. 2009: 288-296.

[30] NEWMAN D, LAU J H, GRIESER K, et al. Automatic evaluation of topic coherence[C]//The 2010 annual conference of the North American chapter of the Association for Computational Linguistics. Los Angeles: Association for Computational Linguistics. 2010: 100-108.

[31] Huang Jiajia, Li Pengwei, Peng Min, et al. Research on deep learning-based topic models[J]. Chinese Journal of Computers, 2020, 43(5): 827-855.

[32] RODER M, BOTH A, HINNEBURG A. Exploring the space of topic coherence measures[C]//Proceedings of the eighth ACM international conference on Web search and data mining. New York: Association for Computing Machinery, 2015: 399-408.

[33] YOKOYAMA S, SANADA H. Logistic regression model for predicting language change[A]//KOHLER R. Issues in quantitative linguistics. Lüdenscheid: RAM-Verlag, 2009.

[34] THORLEUCHTER D, POEL D. Weak signal identification with semantic Web mining[J]. Expert systems with applications, 2013, 40(12): 4978-4985.

[35] CHUANG J, MANNING C D, HEER J. Termite: Visualization techniques for assessing textual topic models[C]//Proceedings of the international working conference on advanced visual interfaces. New York: Association for Computing Machinery, 2012: 74-77.

[36] SIEVERT C, SHIRLEY K. LDAvis: A method for visualizing and interpreting topics[C]//Proceedings of the workshop on interactive language learning, visualization, and interfaces. Baltimore: Association for Computational Linguistics, 2014: 63-70.

[37] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C/OL]. NAACL-HLT, 2019(1). [2021-05-25]. https://arxiv.org/abs/1810.04805.

[38] FRANKLAND R, SMITH A D, SHARPE J, et al. Calibration of topic models with overlapping data[J]. British actuarial journal, 2019(24). [2021-06-25]. http://dx.doi.org/10.1017/S1357321719000151.

[39] ELAKROUCHI M, BENBRAHIM H, KASSOU I. Early warning signs detection in competitive intelligence[C]//The 25th International Business Information Management Association conference. Amsterdam: Association for Computing Machinery, 2015: 512-523.

[40] BLANCO S, LESCA H. Business intelligence: Integrating knowledge into selection of early warning signals[EB/OL]. [2021-06-25]. http://veille-strategique.eolas-services.com.

Author Contributions: Yang Bo: Paper revision and review Shao Wanting: Model construction and paper writing

Note: Figure translations are in progress. See original paper for figures.