Knowledge-Based Systems 212 (2021) 106650

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

End-to-end LDA-based automatic weak signal detection in web news

Manal El Akrouchi *, Houda Benbrahim, Ismail Kassou

IRDA, Rabat IT Center, ENSIAS, Mohammed V University in Rabat, Morocco

ARTICLE INFO

Article history:
Received 1 July 2020
Received in revised form 25 November 2020
Accepted 2 December 2020
Available online 4 December 2020

Keywords:
Weak signals
Topic modeling
Latent Dirichlet allocation

ABSTRACT

An extremely competitive business environment requires every company to monitor its competitors and anticipate future opportunities and risks, creating a dire need for competitive intelligence. In response to this need, foresight study became a prominent field, especially the concept of weak signal detection. This research area has been widely studied for its utility, but it is limited by the need of human expert judgments on these signals. Moreover, the increase in the volume of information on the Internet through blogs and web news has made the detection process difficult, which has created a need for automation. Recent studies have attempted topic modeling techniques, specifically latent Dirichlet allocation (LDA), for automating the weak signal detection process; however, these approaches do not cover all parts of the process. In this study, we propose a fully automatic LDA-based weak signal detection method, consisting of two filtering functions: the weakness function aimed at filtering topics, which potentially contains weak signals, and the potential warning function, which helps to extract only early warning signs from the previously filtered topics. We took this approach with a famous daily web news dataset, and we could detect the risk of the COVID19 pandemic at an early stage.

1. Introduction

Along with the growth of businesses, competition among companies has become intense, and obtaining valuable information of the competitors has become necessary. Thus, gathering the information of competitors is the basic idea behind competitive intelligence (CI). Many aspects of competitive intelligence, such as opinion mining and foresight studies are being used, and the process of obtaining such intelligence differs based on the company's needs [1].

Setting anticipative actions as a goal, companies try their best using foresight techniques and designing strategic tools like Scenario Planning and Horizon Scanning [2,3] to achieve that purpose. This issue has incited business decision-makers to gather any available information about their competitors and analyze it in anticipation of future developments and changes, leading to weak signal detection.

Detecting weak signals and anticipating future scenarios have been the goal of many researchers; hence, many techniques like multi-word analysis [4,5] are used to get the most insights from a word or a document. Besides, text-mining experts are increasingly interested in this field [4,6–8]; However, most of them needed human experts to aid detection [9]. Nonetheless, despite the demand for more automation, a considerable number of studies have used various methods that still require the involvement of a human expert.

The increase in the amount of text data on the Internet has urged scholars to fully utilize it for foresight analysis; Moreover, understanding and manipulating language are too complicated. For this reason, improvements in processing power and natural language processing (NLP), supporting the extraction of insights from text data [10] have been growing exponentially, covering many fields from healthcare to financial industries, including weak signal detection. Therefore, different techniques like Deep Learning and Neural Networks methods are used and proposed to handle different challenges.

Deep learning techniques and algorithms like convolutional neural networks (CNN) and recurrent neural networks (RNN) made remarkable advancements in the NLP field. CNNs are very useful in understanding the semantics and mining semantic signs in a sentence, and RNNs, on the other hand, resolve dependencies in language and similar sequence modeling tasks [11]. As an example of those techniques, Capsule Networks (CN) proposed by [12]. Initially designed for image classification [13], CN are nowadays promisingly used in many NLP applications like Text Classification [14,15] Intent Detection in question-answering systems [16], Relation Extraction [17], Machine Translation [18], etc....

* Corresponding author.
E-mail addresses: manal.elakrouchi@gmail.com (M. El Akrouchi), houda.benbrahim@um5.ac.ma (H. Benbrahim), ismail.kassou@um5.ac.ma (I. Kassou).

https://doi.org/10.1016/j.knosys.2020.106650
0950-7051/

Many other applications of these techniques were proposed in recent works. In the context of Multi-label Text Categorization, for example, the authors of [19] proposed a CNN and RNN hybrid method capable of efficiently representing text features and modeling high-order label correlation. Another application example is Text Generation; the paper [20] ensembled reinforcement learning, generative adversarial networks (GANs), and RNNs to improve the model's generalization capability and learn the structure of sentences directly with RNNs. A final example is Word Dependencies Learning; In [21], the authors introduced a variable-order belief network framework to learn the word dependencies, using dynamic Gaussian Bayesian networks and deep belief networks to predict and classify dynamic multivariate Gaussians.

These improvements in NLP techniques like word embeddings have shown outstanding contribution in capturing similarity between words and predicting a word based on its context [11]. Another one proposed was Embedded Topic Model [22]; the authors proposed a generative model that joins regular topic models with word embeddings. However, these techniques give better results when applied to labeled data compared to unlabeled data [11,22]. Moreover, in the context of weak signals detection, text data is generally unlabeled, like the case of web articles. Hence, the use of deep learning based NLP techniques does not ensure the full automation of the weak signals detection process.

However, traditional topic modeling techniques showed the ability to ensure full automation and have attracted many researchers using old, and new methods [23]. That is why we use a well-known and widely used technique in topic modeling, which is Latent Dirichlet Allocation (LDA) [24].

LDA is an unsupervised machine learning technique that operates independently on human input or intervention to determine topics. Many works used LDA to detect weak signals: The work proposed by [25] has focused on detecting weak signals using dynamic LDA: based on tweets, they used the LDA algorithm over time to extract topics and detected weak signals using the Sankey diagram visualization of topic evolution. In [26], the authors used Dynamic Topic Modeling to track weak signals' life cycle over time.

Consequently, in this study, to fully automate the weak signals detection process, we propose a new method to filter both topics and terms generated using LDA and only extract early warning signs. The main goal is to detect words that can hide important and significant information and can be qualified as weak signals. We propose new functions of filtering topics and terms to obtain early warning signs. The proposed system is fully automatic and does not rely on any initial keywords; it was applied to a dataset of web news to test the detection of weak signals related to COVID19.

The paper is structured as follows: we will first present some definitions of weak signals and LDA and explain their mechanism. The third section presents the proposed automatic early warning signs detection system: we will explain the full process, including the functions we defined to assess and filter topics and terms. The subsequent section describes the case study applied to the New York Times web news to detect early indicators of the COVID19 pandemic. Finally, we exhibit the results of this application study and explain our findings.

2. Methodological foundations

2.1. Weak signals

A large number of studies show diverse definitions of weak signal terminology. Igor Ansoff [27], one of the first contributors to the study of weak signals in 1975, defined them as "Symptoms of a possible change in the future" [28]. According to him [29], for an organization to respond quickly to an uncertain environment, it should be prepared ahead of time to respond to any signs of information about possible threats and opportunities. He defined weak signals as "external or internal warnings that are incomplete to permit an accurate estimation of their impact and/or to determine a complete response" [28–30].

In traditional foresight studies, including weak signal detection methods, the scanning of early warning signs has relied heavily upon the intuitive interpretation of human experts, which is costly and time-consuming; however, the results also vary based on the analyzer's perspectives.

In the big data era, as the size of available information increases exponentially, it is becoming more challenging for experts to effectively scan topics from various information sources. Therefore, many scholars have been using text mining techniques such as information retrieval to extract information sources from the Internet, focusing mainly on web data such as social media and academic papers [6,31–36].

Many researchers have focused on automating data collection techniques. Consequently, keyword filtering in terms of weakness is reliant on the manual efforts of experts. Generally, this process of detection can be semi-automatic, especially when analyzing data based on keywords provided by experts. In [34], parts of the process were performed manually, whereas [6] used keyword-based mining techniques.

Some scholars have attempted to overcome this shortcoming by implementing a fully automatic process. Research on full automation of weak signal detection is still in its initial stages, and this is shown by the low number of papers and projects in this field. Gutsche [26] presented an automatic process of weak signal detection and forecasting using temporal web mining, dynamic topic modeling, and time series analysis. This study follows the same approach as full automation, except for the idea that a topic can be viewed as a weak signal. In this study, we perform deep filtering on both topics and terms to obtain better results with weak signals.

2.2. Latent Dirichlet allocation

Topic models have received particular interest as text data have been increasing extensively. They are used to find latent and hidden information from a text corpus. An aspect of topic modeling is the abstraction of a set of words that are semantically close to each other. Topic models are mainly built using statistical methods, and many techniques have been defined, such as latent semantic analysis (LSA), latent semantic indexing (LSI), and LDA proposed by [24].

A recent research trend in NLP is the word embeddings. Word embeddings are a powerful approach for modeling semantic relations between words [37]. Combined with other NLP techniques such as topic modeling, they formed a new approach to the semantics of topics, and their evolution in time, resulting in techniques like Topic Dynamic Word Embeddings [37,38] and Dynamic Embedded Topic Model [39].

LDA is very famous for discovering latent topics, and many variations of it have been developed to satisfy research projects' needs. Dynamic LDA was proposed by [40]; it uses time series to capture the evolution of topics through time. Concept-LDA was presented by [41], which combines LDA with concepts and named entities. Given the importance of semantics of words, semantic similarity based on LDA has attracted many researchers' attention. In [42], the authors investigated semantic similarity measures at both word and sentence level based on LDA and LSA. In [43], the authors proposed a semantic variation of LDA named Sentic-LDA; it integrates common-sense knowledge-based

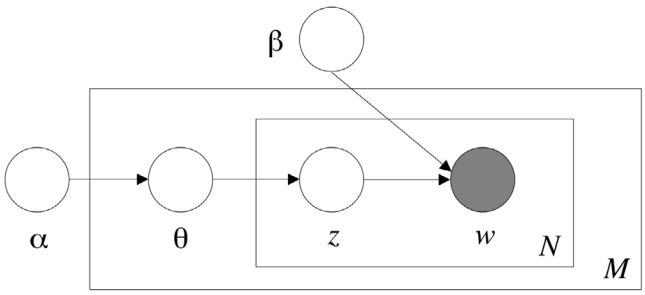

Fig. 1. The graphical model representation of LDA.

similarity and gives more semantics to words than syntax in aspect-based sentiment analysis.

Regarding weak signals detection, LDA is widely used to detect hidden information, potentially referring to weak signals. In [44], the authors proposed LDA augmented with Word2Vec that detects a topic related to a weak signal characterized by low frequency (few words per document) and low visibility (few documents having this signal). In [45], the authors proposed a method to automatically assess topics produced by LDA, helping in detecting weak signals.

LDA provides a generative probabilistic model [24] that describes how the documents in a dataset are created. This method is capable of identifying semantic topics from a collection of text documents in an unsupervised manner.

The input of LDA is a collection of documents and two parameters, α and β. The output is a probabilistic model that describes [46]:

- The number of words that belong to the topics.
- The relevance of the topics with the documents.

The LDA model is illustrated as a "probabilistic graphical model" [24] in Fig. 1. The boxes are "plates" representing replicates. The outer "plate" represents the documents, while the inner "plate" represents the repeated selection of topics and words within a document. α and β are two hyperparameters of LDA.

To explain the general process of LDA, we define the following terms [24]:

- A word is an item from a vocabulary indexed by {1, …, V}.
- A document is a sequence of N words denoted by $W = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.
- A corpus is a collection of M documents denoted by $D = \{W_1, W_2, \ldots, W_M\}$, where W_m is the mth document.

According to [24], LDA follows generative steps for each document W in a corpus D for a pre-defined number of topics k, as presented in the following process:

1. Choose $\theta \sim \mathrm{Dir}(\alpha)$
2. For each of the N words w_n:
   - Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
   - Choose a word w_n from $\mathrm{p}(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Finally, the joint probability of the model can be defined as follows:

$$\mathrm{p}(\mathrm{W}, \mathrm{Z}, \theta, \phi \mid \alpha, \beta) =$$

$$\prod_{z=1}^{k} p(\phi_z \mid \beta) \prod_{d \in D} p(\theta_d \mid \alpha) \prod_{j=1}^{|d|} p(w_{d,j} \mid \phi_{z_{d,j}}) p(z_{d,j} \mid \theta_d)$$

where $W = \{w_{d,j}\}_{d,j}$ is a word set, $Z = \{z_{d,j}\}_{d,j}$ is the topic assigned to each word, $\theta = \{\theta_d\}_d$ is a set of per-document topic combinations, $\phi = \{\phi_z\}_z$ is a set of per-topic term combinations. α and β are the hyperparameters of the Dirichlet priors.

When using LDA in topic modeling, one of the challenges is to specify the most accurate number of topics k. One elementary way to choose k is to run the LDA model with different values of k and then fix the one that gives the most significant coherence. In the next section, we will describe the choice of the number of topics.

3. Design of the proposed method for automatic detection of early warning signals

3.1. Overview

The process of detecting early warning signs poses challenges to researchers at every step. Many projects have been proposed to detect weak signals, but only a few of them are fully automatic. There is a noticeable lack of automation in this field. Most studies rely on manual inputs or expert opinions [23].

Automated approaches include statistical, linguistic, and semantic techniques from natural language processing (NLP), such as generalized sequential patterns [47], named entity recognition [48], Point-of-Speech (PoS) tagging [49–51], SAO structures [52,53], and latent semantic analysis [54]. When using natural language processing, these techniques are mainly employed to tag and filter extracted features. Text data is then transformed into a structured vector space model, and mining techniques are applied.

Generally, the process of weak signal detection starts with some initial keywords given by human experts. To overcome human expert intervention, we would like to extract terms that may be weak signals, without the aid of keywords. Thus, in this study, we propose an early warning sign detection system to automatically and deeply filter topics and terms through two proposed functions: "Weakness" function, which helps to filter topics and keep only those that potentially contain weak signals; "Potential Warning" function, which helps to mine through previously filtered topics and perform further filtering of terms to retain only weak signals.

The main advantages of this proposed model are as follows. Generalization: extracted weak signals are not specific to a certain field or subject, but are warning signs that should be considered in the chosen period of time, and decision-makers are responsible for choosing which signal is more relevant to their needs; automation: the proposed method automatically detects weak signals from full text web news without human expert intervention.

The proposed method is illustrated in Fig. 2. In this study, we used web content; therefore, the first step was data collection through a period of time. The collected data served as input for the foresight study. In the first part of the foresight study, i.e., the analysis part, the data must be subjected to several preprocessing procedures. The second and most important part is the filtering step. Filtering aims to find potential topics that contain some weak signals. This step is derived through the identification of topics by applying a topic modeling LDA algorithm and then applying filtering functions that reduce the number of topics. In the final part of the foresight study, we use another filtering method based on the rarity of words and other parameters to extract only potential warning signs. The final step in the output enhances the detected weak signals through the use of word embedding.

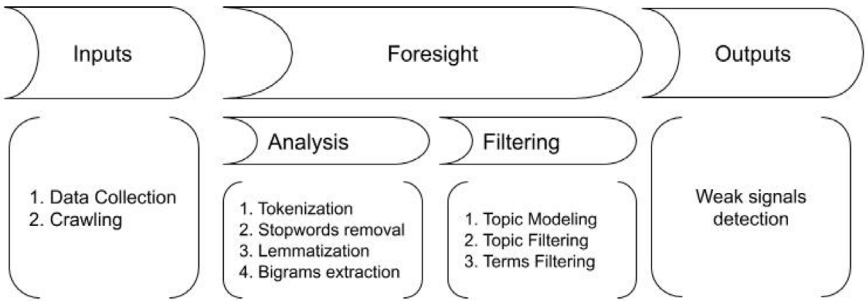

Fig. 2. Process of detecting early warning signs.

3.2. Data collection and pre-processing

The Internet has now become the primary source of information, the size of which is exponentially increasing over time. Many sources are currently available for web news, such as web journals, blogs, and social media. The most reliable source of structured data is web journal news, which was our source of data in this study.

To use the data for the proposed system, it must be processed further. The goal of the preprocessing procedure is to filter irrelevant terms within the corpus and prepare the data for the next step. This procedure starts with tokenization. Tokens are then processed to remove the common stop words in the English language. This is followed by the lemmatization step, in which we mainly keep nouns, verbs, and adjectives using part-of-speech tagging. Finally, the last step of the pre-processing procedure is to extract bigrams, and they are grouped as new terms. Bigrams are the concatenation of two consecutive tokens or words, and they are helpful in enriching the word corpus with commonly used bigram words. After this step, all documents in the corpus are prepared for the next task.

3.3. Early warning sign detection system

To design a fully automatic early warning sign detection system, we considered using unsupervised text mining techniques related to topic modeling. One of the best methods for this is LDA. LDA is generally used to extract the trendier topics from a text dataset. In our approach, we look for the rarest topics that can potentially lead to weak signals. This approach is adopted more often to detect weak signals [23,25,34].

In contrast to studies that depend on keywords for weak signal detection [6], topic models consider aspects of meaning rather than words. This study follows the conventions of applying topic models for weak signal detection, but without accepting that all topics are "containers" of weak signals. We believe that once terms are gathered for a topic, the weakness of these terms depends on the weakness of their topic. For this reason, we propose a method that first filters out the topics. The resulting topics are then subject to another filtering method to extract only potentially weak signals.

3.3.1. Topic model training

One of the main challenges of LDA is finding the optimal value of the number of topics **k**. The values of hyperparameters α and β represent document-topic density and word-topic density, respectively. They play an important role in establishing coherence between topics and terms.

To find the optimal number of topics, researchers have proposed evaluating topic models based on the topic coherence for different values of k [55]. Recently, coherence measures [56,57] were proposed to weigh topic quality that corresponds more to human interpretability. Our study follows this approach and applies the topic coherence measure c_v proposed by [58].

To find the model with the highest coherence, we ran a set of tests sequentially, changing one parameter value at a time, keeping others constant and running them over the corpus set. We use c_v as the choice of coherence metric for achievement testing. This metric [58] is based on a sliding window, one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. The resulting model of k number of topics returning the highest coherence is then chosen as the best model.

3.3.2. Topic filtering

In this subsection, we propose a new method of topic filtering that helps in assessing the weakness of the topic. This method is derived from a logistic function.

Generally, the logistic function is used to illustrate the progress and growth of a population, an innovation, and other parameters. In linguistics [59], it is used to model language change; a term that is marginal initially begins to spread more quickly with time, but if it is weak, it remains marginal. The same principle applies for detecting weak signals. Using the logistic function, we are more interested in low values because they represent topics and rarity of terms.

We define the "Weakness" method as a function based on three main measures: closeness centrality, weight of topic t based on its coherence within the model, and the autocorrelation function.

The first measure, closeness centrality, is based on the distances between a topic t and the other topics. The distance between the two topics describes the similarity of these topics. Many distance measures can be used to calculate the similarity, such as Jaccard distance, cosine-distance, and Hellinger distance. Scholars have studied several similarity measures to find the most efficient filter [25]. Their judgment was that a distance measure must present a sigmoid-based variation [25]. In this study, we follow this approach and choose the Hellinger distance. Based on this distance, we calculate the closeness centrality CC as defined in function (1):

$$CC(t) = \frac{1}{\sum_i d(t, t_i)} \tag{1}$$

where d is the Hellinger distance.

We now calculate the second measure: Weight of topic. This measure is based on the coherence of the relevant topic within the model. The weight of a topic is the value assigned based on the significance of each topic. We define the weight of topic W, the coherence of a topic t by the total coherence of all topics, as in function (2):

$$W(t) = \frac{Coh(t)}{\sum Coh} \tag{2}$$

where Coh(t) is the coherence score of topic t.

In economics, "trend analysis" usually refers to the analysis of past trends in market data; it allows one to predict what might happen to the market in the future. It quantifies and explains trends and patterns in "noisy" data over time. Various tools are used to analyze trends in data. They range from relatively simple methods, such as linear regression, to more complex tools, such as the Mann–Kendall test, which may be used to search for non-linear trends. Other popular tools include autocorrelation analysis.

Autocorrelation describes the relationship between observations of the same variable over different periods of time. It is a type of serial dependence. Specifically, autocorrelation is when a time series is linearly related to a lagged version of itself. In our case, the document frequency of a topic may change over time. For this, autocorrelation of each topic over a number of days of the month helps us filter out topics that may not contain weak signals. The autocorrelation AC is defined as follows (3):

$$AC(t) = \frac{Cov(t)_k}{Var(t)} \tag{3}$$

where $Cov(t)_k$ is the covariance of topic t at lag k.

The three functions mentioned above, (1), (2), and (3), are used to form the Weakness function WK. We define the Weakness function of a topic t as follows (4):

$$WK(t) = \frac{W(t) * CC(t)}{1 + \exp^{-(AC(t))}} \tag{4}$$

After measuring the WK function, we narrow down the results to obtain only those topics that potentially have weak signals. Based on the definition of weak signals, rarity is the main characteristic of their weakness, and their movement over time will be slow. In other words, the lower a topic's WK function value, the weaker its corresponding signal is. Therefore, based on the WK function, only topics with low values are deemed to be considered weak topics.

In most cases, weak signals form no more than 20% of information (such as the Pareto principle). In addition, as in most information and data, noise can generally vary between 0% and 5%. For example, in the case of [60], thresholds were used based on human expert opinion in their model, and it was decided that the noise threshold has a value of 0% to 2%, which represents rarely occurring words that do not carry meaningful information. Weak signal thresholds have values from 2% to 6%.

In our case, the design of the two functions leads to the selection of very low values. We believe that having the lowest WK value does not necessarily guarantee a weak signal; it could be noise. Thus, we proposed new values of thresholds: under 1% for noise (that may in the future transform to weak signals) and under 10% for weak signals.

Fig. 3 illustrates the distribution of signals, from noise to strong signals.

Therefore, we decided to disregard the lowest WK values and used two percentiles for the WK function values (representing weakness thresholds) to obtain only the topics indicating WK results below 10% and above 1%, omitting the "noise".

Finally, Algorithm 1 summarizes the steps of the topic filtering process presented above.

The filtered topics obtained from this process serve as an input to the next step of term filtering.

3.3.3. Term filtering

For topic filtering, we only extracted those topics that possibly contained weak signals. On the other hand, obtaining the filtered topics does not necessarily imply that all the related terms are weak signals. Term probability within a topic and frequency, for example, can make a difference between terms. In addition, the number of terms that we should extract from these topics remains a challenge.

We know that the topics generated by LDA are not always highly explicable by humans. Many studies have been proposed to resolve this challenge and find useful terms for more interpretability of the topics.

In [61], two measures of the usefulness of terms were introduced to obtain intelligible topics: "distinctiveness" and "saliency". These quantities measure the information conveyed by a term regarding a certain topic. The Kullback–Liebler divergence between the likelihood that an observed word is generated by a latent topic and the marginal probability of a topic generates distinctiveness. Saliency is the product of the overall frequency and distinctiveness of the term.

As in [62], which involved calculating the relevance of terms for different topics and deriving the most relevant terms, we propose that the measure of relevance as a superior indicator for topic interpretation over term probability.

When it comes to weak signal detection, finding the most relevant term does not guarantee a corresponding weak signal. Therefore, this study follows a new method, derived from a logistic function, and is based on the probability and frequency

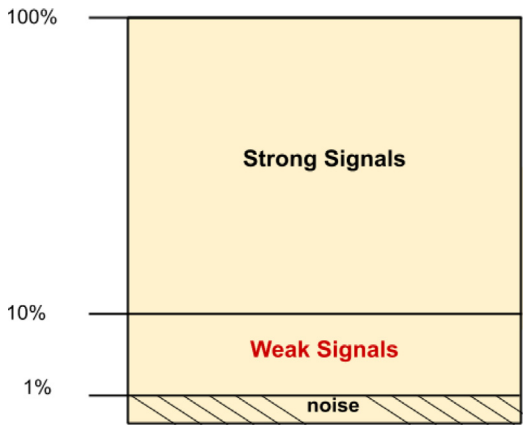

Fig. 3. Diagram of signal distribution.

Algorithm 1: Topic Filtering Process

Input:
- Corpus of a given period ;
- LDA model ;

Output: Filtered_Topics

```
begin
    for each topic t do
        Calculate CC(t) ;
        Calculate W(t) ;
        Calculate AC(t) ;
        Calculate WK(t) ;
    end
    Calculate P_10 : percentile of 10% ;
    Calculate P_1 : percentile of 1% ;
    for each topic t do
        if P_1 < WK(t) < P_10 then
            Filtered_Topics ← t ;
        end
    end
end
```

of the term within documents that are related to the topic corresponding to the term.

We define the "potential warning " (PW) function of a term w as follows (5):

$$PW(w) = \frac{NF(w)}{1 + \exp^{-(\phi(w) * \log(\phi(w)))}} \quad (5)$$

where $NF(w)$ is the normalized frequency of a term w in a topic t, and $\phi(w)$ is the probability of the term w in the topic t obtained with the LDA model.

As mentioned before, the point of weak signals is the rarity, and the terms to be qualified as weak signals should not have extremely low values; otherwise, we will have meaningless terms. Based on this, we use the percentile to extract only the terms with values above 1% and below 10%.

Algorithm 2 explains the process of term filtering.

Algorithm 2: Term filtering process

Input:
- minimum document frequency: p ;
- Filtered_Topics: containing terms and probabilities ϕ ;
- Frequency: Number of documents related to topic t in Filtered Topics, containing a term w;

Output: Filtered_Terms
begin
  for *each topic t in Filtered_Topics* **do**
    if *Frequency* > *p* **then**
      foreach *term w* **do**
        Calculate NF(w) ;
        Calculate PW(w) ;
      end
      Calculate P_10 : percentile of 10% ;
      Calculate P_1 : percentile of 1% ;
      foreach *term w* **do**
        if $P\_1 < PW(w) < P\_10$ **then**
          $Filtered\_Terms \leftarrow w$;
        end
      end
    end
  end
end

To further obtain words related to these weak signals for more interpretability, we use the Word2Vec technique, which we will explain in the next section.

3.3.4. Early warning sign output

After the extraction and detection of weak signals using the two functions defined above, the results obtained must be analyzed for a better understanding. This could be achieved if we have more words similar to the signals obtained; in that case, we can understand the contexts of these signals. To that end, we use Word2Vec, a neural network-based model, representing words in the corpus as a vector with contextual concepts [63]. The general idea of this model is to represent each word of the training set as a real vector that captures the contexts in which this word appears. The network is trained to reconstruct the context of the words, so that the words that share a common context in the corpus are close to each other in the semantic space.

Word2vec provides two model architectures: the Continuous Bag-Of-Word model (CBOW) [63] and the skip-gram model [64]. In the CBOW model, the context of the current word is predicted based on the context of the neighboring words, while the skip-gram model predicts the context based on the current word instead of [65]. Fig. 4 from [64] presents a graphical representation of the CBOW and Skip-gram models.

Depending on their research need, some researchers prefer using the Google pre-trained Word2Vec model [63,66] as in [67]. Other researchers prefer using contextual word vectors constructed from a domain-specific text corpus like in [68]: the authors investigated the performance and robustness of contextual Word2Vec vs. generic Word2Vec in classifying unstructured text data to specify Zika and Ebola outbreaks. Our study followed this contextual approach, so we built the Word2Vec models using each filtered topic's documents. The resulting models would be used to highlight words semantically related to the previously obtained weak signals. Consequently, the proposed system's output is a list of words similar to each weak signal.

Finally, the following Algorithm 3 describes the use of Word2Vec to generate terms similar to the weak signals.

Algorithm 3: Word2Vec: similar terms generation

Input:
- Filtered_Terms;
- Corpus of documents related to topic t from Filtered_Topics;

Output: Similar_Words
begin
  foreach *topic t in Filtered_Topics* **do**
    $W2V\_Model \leftarrow Train\ Word2Vec\ Model\ (Corpus(t))$;
    foreach *term w in Filtered_Terms* **do**
      $Similar\_Words \leftarrow W2V\_Model.Most\_Similar(w)$;
    end
  end
end

The output of this algorithm marks the end of the proposed weak signal detection system. In the next section, we present an application of this system.

4. Application study

Identifying early warning signs is becoming the ultimate goal of businesses; thus, weak signal detection has gained an important place in competitive intelligence. In this study, we test our proposed system on a dataset covering the COVID19 pandemic. Being able to detect early warning signs of a pandemic would be of great help to decision-makers.

For this reason, we apply the proposed system to web news reported by The New York Times to detect weak signals related to COVID19. We know that this virus was not known in December 2019. The news started covering the matter since January 2020, and it became a strong signal around February 2020. Therefore, any weak signal pointing to this pandemic would be helpful. Thus, we try to detect weak signals in December 2019 and test their evolution over time in January and February 2020.

4.1. Application environment

The application of this study is tested using the tool Gensim[1] [69], a Python open-source library for unsupervised topic modeling and natural language processing, which uses modern statistical machine learning. We use this tool for all processes of the proposed system.

[1] https://radimrehurek.com/gensim/

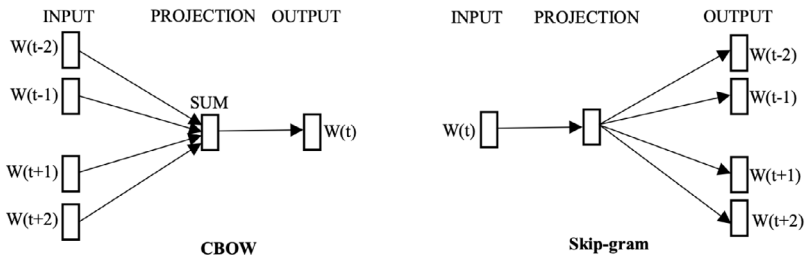

Fig. 4. Architectures of CBOW and Skip-gram models.

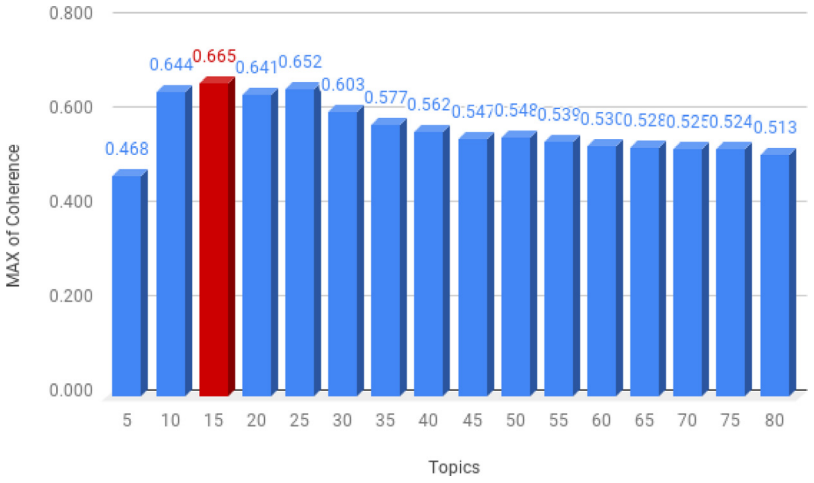

Fig. 5. Histogram of Coherence.

4.2. Data collection and pre-processing

The data chosen for this analysis were online web news. Many sources are now available for web news, but to have more reliable data, we need to collect news from a trustworthy source that has articles of all domains and areas. For this, we chose The New York Times.

Our goal for this test was to detect weak signals related to the COVID19 pandemic. Hence, we chose data for three months: December 2019 to February 2020, as the presence of news related to the virus was less during this period.

Using the New York Times API to search for archived articles, we tried to collect all the news published during this period, regardless of any query or keyword to have more hidden information that can potentially give rise to weak signals. The lesser the number of queries, more is the extent of automation.

The number of entries for December 2019 were 5659, 7585 for January 2020, and 6398 for February 2020. To retrieve the full text of these documents, we used a web crawler, after which the extracted data is cleaned to obtain only the text and date of publication. After cleaning the dataset, the final number of results was 5522 for December 2019, 6364 for January 2020, and 5923 for February 2020. The output of the data collection step is a dataset of dateable full-text documents from the three highlighted months.

The pre-processing procedure, as stated before, has 4 steps. Using Gensim, we tokenize the text, stop words are then removed, terms are lemmatized, and bigrams are extracted and grouped as new terms throughout the corpus. The web documents in the corpus were prepared for further analysis.

Table 1
Example of values proposed to hyperparameters Alpha α and Beta β.

Hyperparameter	Values
Alpha α	0.01 0.31 0.61 0.91 symmetric asymmetric
Beta β	0.01 0.31 0.61 0.91 symmetric

4.3. Application of the early warning sign detection system

4.3.1. Topic modeling

To find the best topic model, we trained multiple models for different values of the number of topics k, and the hyperparameters α and β. Their coherence measure c_v was evaluated. Gensim has a module called "LdaModel". This module allows both LDA model estimation from a training corpus and the inference of topic distribution in documents. Therefore, we began by first modeling the data of December 2019.

Finding the best values of parameters for the model is crucial in topic modeling [70]; thus, we set a range of values for the number of topics k from 5 to 85. As for hyperparameters α and β (named "Eta"in gensim), the possible priors can be symmetric or asymmetric [71,72]. A symmetric distribution means that each topic is evenly distributed throughout the document, while an asymmetric distribution favors certain topics over others. Table 1 shows the range of values proposed.

According to Fig. 5, the best coherence corresponds to the number of topics k value of 15. Going back to the table of different values of hyperparameters α and β, we obtain the following values:

- Alpha α = asymmetric
- Beta β = 0.61

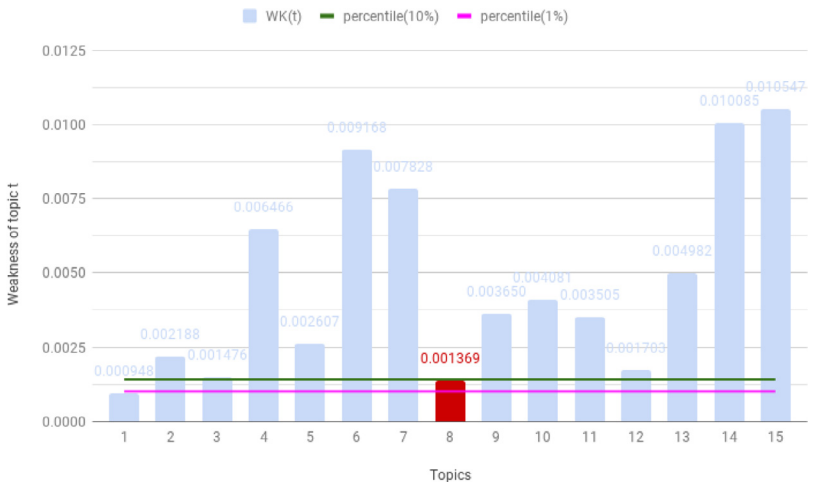

Fig. 6. Topic Filtering of December 2019 data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

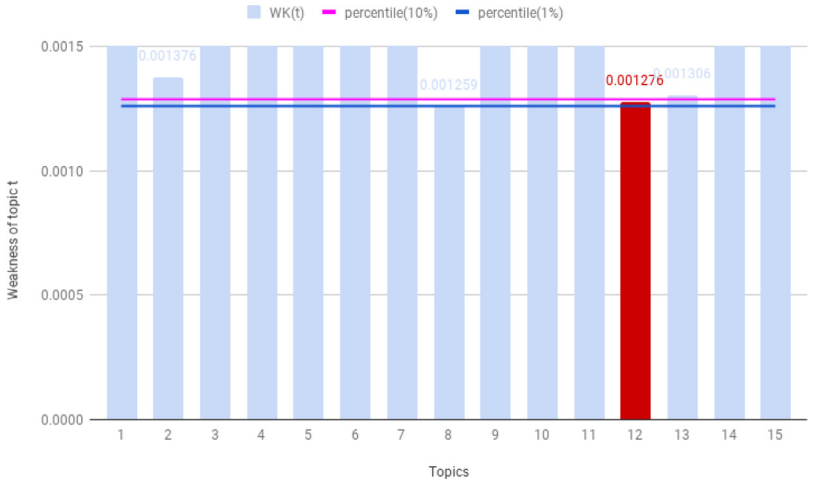

Fig. 7. Topic Filtering of January 2020 data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

With these parameter values, we trained the LDA model on the data of each month separately. The resulting models would then be ready for the subsequent filtering steps.

4.3.2. Topic filtering

In this section, we apply the algorithm presented above to filter the topics. We first calculate the distance of each topic with other topics using the Hellinger distance. The obtained matrix (15∗15) is then used to measure the first function: Closeness Centrality.

We then measure the weight function W(t). Since it is based on the coherence of each topic, we use Gensim's method to measure the coherence of each topic. We then assign to each topic its value, and finally measure W(t). The last function, autocorrelation AC(t), is calculated based on the frequency of documents of all topics, for each day. Table 2 provides an example of this frequency.

After counting this frequency, we calculate the autocorrelation function AC(t). As previously mentioned, the autocorrelation takes a lag k as a parameter. A key issue in trend estimation is that there is no sample overlap between the observations used to calculate the autocorrelation; thus, it is beneficial to examine changes over the period with longer lags [73]. In many cases, non-overlapping time-series is less autocorrelated than overlapped data [74]. The more non-overlapped the data are, the lower the autocorrelation will be.

Table 2
Example of document frequency for topics T1 to T15, for the first week of December 2019.

Date	T1	T2	T3	...	T12	T13	T14	T15
2019-12-01	52	51	50	...	48	40	0	0
2019-12-02	119	120	112	...	111	84	0	0
2019-12-03	313	298	294	...	290	242	4	0
2019-12-04	182	177	172	...	161	142	0	0
2019-12-05	349	341	336	...	319	282	3	4
2019-12-06	175	175	160	...	164	152	1	0
2019-12-07	189	187	182	...	177	154	0	0

In our case for weak signal detection, we want to minimize the WK function, which is a minimization of the AC function. This can be achieved by reducing the overlap period between the time series (higher lag value). Therefore, the best time lag that we can choose for the calculation of the AC function will be half of the chosen period. In this test, the value of time lag k is taken to be 15 (days). Using this time lag value, we calculate the AC function.

The past calculations help us calculate the "Weakness" function WK(t), and the percentile needed, as explained before. Figs. 6, 7, and 8 illustrate the results for December 2019, January 2020, and February 2020, respectively. The topics marked in red in both

Table 3
Frequency of documents for each filtered topic of December 2019, January 2020, and February 2020.

Month	Topic	Total Docs
December 2019	T8	5119
January 2020	T12	2410
February 2020	T4	2457

Table 4
Example document frequency for each term in the filtered topic T12 from January 2020.

Term	Frequency of documents
government	0.53341
flight	0.13852
report	0.31146
call	0.44350
health	0.22600
travel	0.13973
plane	0.15998

Table 5
Filtered terms for topic T8 from December 2019.

Topic	Filtered terms
T8	asset cigarette climate_change debt device **disease** emission employer employment factory farmer housing insurance investor loan manufacturer package payment pension privacy procedure regulator requirement revenue software user wage

Table 6
Terms filtered of topic T12 for January 2020.

Topic	Filtered terms
T12	accident aircraft airline airport cancel crew demonstration european_union evacuate **fever** helicopter **illness** **infection** mask militia pilot regime researcher rocket sovereignty territory **travel_restriction** traveler **vaccine**

the graphs are the filtered topics. These topics have weakness function WK values above the percentile of 1% and below the percentile of 10%.

In December 2019, the topic was T8. For January 2020, we found the topic T12, and topic T4 for February. We then continue to the terms filtering step of these 3 resulting topics.

4.3.3. Terms filtering

Now that we have filtered only 3 topics out of 45 (1 topic per month), we start the terms filtering process. It involves the extraction of only those terms that can be viewed as weak signals. The first step in this process is to obtain the terms of each filtered topic. LDA groups and sorts words in each topic according to its probability in each topic. We need to have a large number of terms in each topic to avoid the possibility of missing any potential weak signals. For this, we obtain 500 terms per topic. These terms are subjected to Algorithm 2 for filtering. The following step consists of counting the frequency of documents for each topic. Table 3 presents the total number of documents for each topic in the data for December, January, and February.

The next step is to count the frequency of documents for each word in a topic. Table 4 presents an example of document frequency for the filtered topic T12 of the January 2020 data.

The frequency helps in the calculation of the "Potential Warning" (PW) function. Using the two percentiles of 1% and 10%, we finally obtain the filtered terms. Tables 5, 6, and 7 present the results obtained for December 2019, January 2020, and February 2020, respectively.

For more interpretability, we use the last Algorithm 3. In this algorithm, the Word2vec model used was CBOW. As mentioned

Table 7
Terms filtered of topic T4 for February 2020.

Topic	Filtered terms
T4	application broker campus climate computer county developer economist emission employer expansion expense facility funding lawsuit license loan payment poverty provider requirement taxis tenant traffic university wage

Table 8
Words similar to the term "disease" from Topic T8 of December 2019.

Filtered term	Similar words
Disease	ailment **infection** tuberculosis mosquito **lung** diagnosis **respiratory** asthma symptom cancer deadly_germ **outbreak** drug_resistant nsaid **illness** coli degenerative resistant anthrax antibiotic **vaccine** cure **diagnosis** infect bacterial **contagious** duodenoscope bacteria **epidemic** pesticide

before, CBOW helps predict the word by context, that is, it maximizes the probability of the target word by analyzing the context. This happens to be a problem for rare words, which in our case, are weak signals.

Using the Word2Vec method, we can obtain the top 30 words most similar to each filtered term obtained in the previous step. The final result is a list of all the filtered terms, with their semantically related words.

After generating words similar to each weak signal that was obtained before, we now move to the discussion section, where we explore the results of the procedures in detail.

4.4. Results and discussion

The results obtained using the proposed automatic weak signal detection system contain all the early warning signs from December 2019 to February 2020; as mentioned earlier, we gathered all the data from that period to obtain more insights. In this study, we were more interested in weak signals related to COVID19.

We start first with weak signals of December 2019. After applying Algorithms 1 and 2 for filtering topics and terms, we can see from Table 5 that there is only one weak signal "disease"that is slightly related to what we are looking for in our study. After applying Algorithm 3 and generating similar words to "disease"(presented by Table 8), we found some important weak signals such as "outbreak", "epidemic"and "lung".

Similarly, when exploring the filtered terms from January 2020 (Table 6), we found that the similar words of the word "disease"obtained from December 2019 in Table 8 started appearing in January 2020, which can be a sign of the evolution of this signal.

Besides, the similar words obtained in December's results evolved in January in two ways:

- As weak signals: like "infection", "vaccine"obtained directly after using the filtering functions (Algorithms 1 and 2).
- As similar words: explored only after applying the Word2Vec model (Algorithm 3) to some weak signals, which is the case for the term "illness" in the filtered topic T12 from January 2020 (Table 9). The words similar to "illness" include some old words such as "epidemic", "disease", "diagnosis"and "outbreak". In addition, we found new weak signals such as "pandemic", "coronavirus"and "sars_epidemic".

Moreover, if we take, for example, the signals "outbreak" and "epidemic", we can infer that there is an early warning sign of an epidemic in December 2019. These signals evolve through time

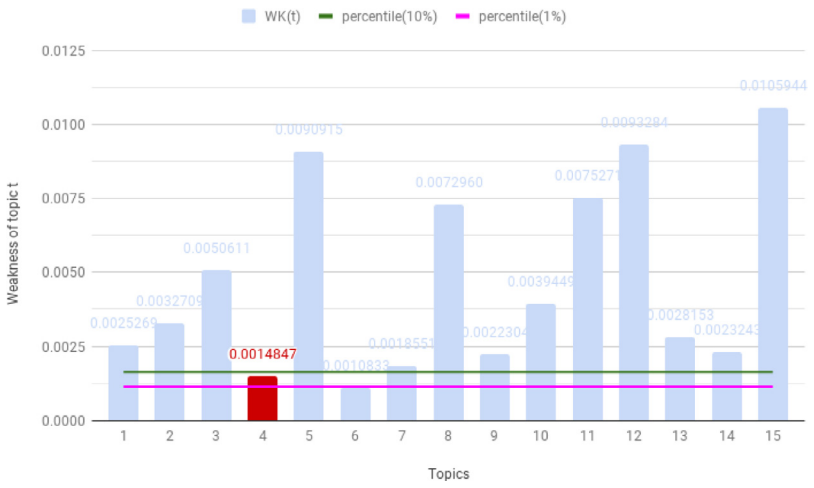

Fig. 8. Topic Filtering of February 2020 data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

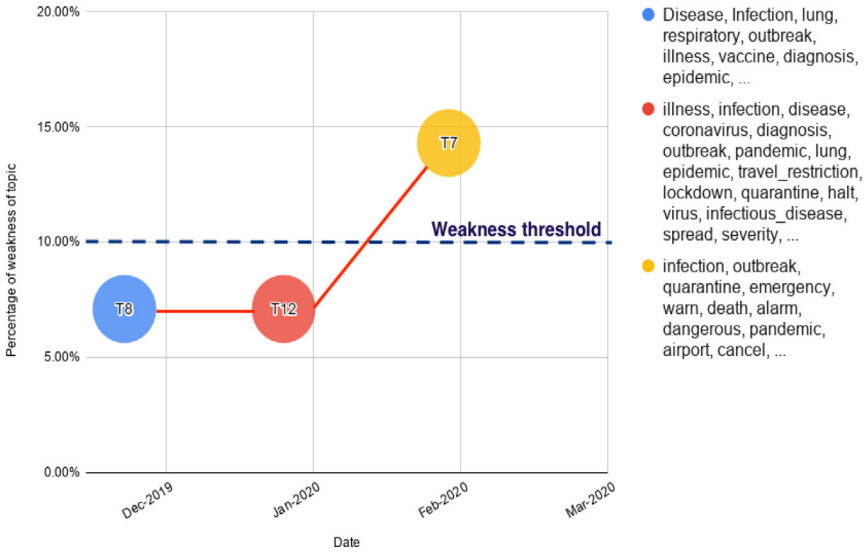

Fig. 9. Evolution of weak signals from December 2019 to February 2020.

Table 9
Words similar to the term "illness" in Topic T12 from January 2020.

Filtered term	Similar words
Illness	mer **infection** flu respiratory influenza **disease** symptom **coronavirus** infect hospitalize**diagnose** pneumonia measle virus pathogen fever **outbreak** colitis **pandemic** transmit transmission **lung** diagnosis sar ulcerative cough **epidemic sars_epidemic** germ human_transmission

and are also detected in January 2020, along with other warning signs such as "pandemic". We also concluded that using contextual Word2Vec to get similar words can enrich the meaning of detected weak signals and detect new weak signals that may grow stronger in time.

As for the results from February 2020, we see from Table 7 that there is no word related to the term "disease". When exploring other topics in the same month, we found that weak signals detected in the previous months are included in T7 with a WK value larger than the percentile of 10%. This finding implies that the signals for December became partially strong, which explains the evolution of the weak signals over time.

Knowing that the pandemic was officially declared in March 2020, we can then deduce from these results that the filtered terms and their similar words can be considered as weak signals.

In Fig. 9, we summarize the results of weak signal evolution from this test. We can see that weak signals detected in December 2019 remained weak in January 2020; but in February 2020, COVID19 was discussed extensively. The signals exceeded the weakness threshold and became partially strong, which explains the detection of new weak signals in February.

5. Conclusion

This study presents a new model for the automatic detection of weak signals based on latent Dirichlet allocation. We proposed two functions for deep filtering. The first function, called Weakness, is used to filter topics created by LDA and consists of three metrics: coherence, to measure the significance of topics; closeness centrality, to measure similarity and closeness between topics; autocorrelation, to measure the evolution of topics over time. These three metrics were combined using the logistic function and percentile. We retained only those topics that had values between 1% and 10% (Fig. 3).

The second function, Potential Warning, is aimed to apply further filtering to the terms of the previously filtered topics. We applied the Algorithm 2 to filter terms and keep only potential weak signals.

These two functions constitute the primary functionality of foresight in the proposed model. The output of this step assists the last step of the process. At the end, for more interpretability, we applied word embedding using Word2Vec models to extract words similar to each filtered weak signal. The final result is a set of early warning signs.

This model was tested to extract weak signals related to the COVID19 pandemic. Using web news from December 2019, we managed to detect early warning signs related to the "epidemic", and we tested the evolution of these signals through January and February 2020.

The purpose of our model is to support decision-makers to automatically detect early warning signs, reducing the time and costs associated with human experts. Future work can improve this system. A large number of examples of weak signal data extracted from the past can improve the choice of thresholds for further insights and anticipation. In addition, the interpretation of the weak signals should be presented to decision-makers, to assist their decisions on which weak signals should be considered. Further studies will include generation of weak signal evolution graphs for better visualization.

CRediT authorship contribution statement

Manal El Akrouchi: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization, Investigation, Writing - review & editing. **Houda Benbrahim:** Conceptualization, Validation, Supervision. **Ismail Kassou:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K. Amarouche, M. El Akrouchi, H. Benbrahim, I. Kassou, Introduction to competitive intelligence: Process, applications and tools, in: Proceedings of the 27th International Business Information Management Association Conference, 2016.

[2] B.L. van Veen, J. Roland Ortt, P. Badke-Schaub, Compensating for perceptual filters in weak signal assessments, Futures 108 (2019) 1–11, http://dx.doi.org/10.1016/j.futures.2019.02.018.

[3] E. Rowe, G. Wright, J. Derbyshire, Enhancing horizon scanning by utilizing pre-developed scenarios: Analysis of current practice and specification of a process improvement to aid the identification of important 'weak signals', Technol. Forecast. Soc. Change 125 (2017) 224–235, http://dx.doi.org/10.1016/j.techfore.2017.08.001.

[4] I. Griol-Barres, S. Milla, A. Cebrián, H. Fan, J. Millet, Detecting weak signals of the future: A system implementation based on text mining and natural language processing, Sustainability 12 (19) (2020) 1–22, http://dx.doi.org/10.3390/su12197848.

[5] I. Griol-Barres, S. Milla, J. Millet, System implementation for the detection of weak signals of the future in heterogeneous documents by text mining and natural language processing techniques, in: ICAART (2), 2019, pp. 631–638.

[6] J. Yoon, Detecting weak signals for long-term business opportunities using text mining of Web news, Expert Syst. Appl. 39 (16) (2012) 12543–12550, http://dx.doi.org/10.1016/j.eswa.2012.04.059.

[7] S.H. Yoo, D. Won, Simulation of weak signals of nanotechnology innovation in complex system, Sustainability 10 (2) (2018) http://dx.doi.org/10.3390/su10020486.

[8] P. Krigsholm, K. Riekkinen, Applying text mining for identifying future signals of land administration, Land 8 (12) (2019) http://dx.doi.org/10.3390/land8120181.

[9] M. El Akrouchi, H. Benbrahim, I. Kassou, Early warning signs detection in competitive intelligence, in: Proceedings of the 25th International Business Information Management Association Conference - Innovation Vision 2020: From Regional Development Sustainability to Global Economic Growth, IBIMA, 2015, pp. 1014–1024.

[10] J. Hirschberg, C.D. Manning, Advances in natural language processing, Science 349 (6245) (2015) 261–266, http://dx.doi.org/10.1126/science.aaa8685.

[11] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing [review article], IEEE Comput. Intell. Mag. 13 (3) (2018) 55–75, http://dx.doi.org/10.1109/MCI.2018.2840738.

[12] G.E. Hinton, A. Krizhevsky, S.D. Wang, Transforming auto-encoders, in: T. Honkela, W. Duch, M. Girolami, S. Kaski (Eds.), Artificial Neural Networks and Machine Learning, ICANN 2011, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 44–51.

[13] M. Kwabena Patrick, A. Felix Adekoya, A. Abra Mighty, B.Y. Edward, Capsule networks – A survey, J. King Saud Univ. Comput. Inf. Sci. (2019) http://dx.doi.org/10.1016/j.jksuci.2019.09.014.

[14] J. Kim, S. Jang, E. Park, S. Choi, Text classification using capsules, Neurocomputing 376 (2020) 214–221, http://dx.doi.org/10.1016/j.neucom.2019.10.033.

[15] W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, Z. Zhao, Investigating capsule networks with dynamic routing for text classification, 2018, arXiv:1804.00538.

[16] C. Xia, C. Zhang, X. Yan, Y. Chang, P.S. Yu, Zero-shot user intent detection via capsule neural networks, 2018, arXiv:1809.00385.

[17] N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, H. Chen, Attention-based capsule networks with dynamic routing for relation extraction, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 986–992, http://dx.doi.org/10.18653/v1/D18-1120.

[18] M. Wang, Towards linear time neural machine translation with capsule networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, Association for Computational Linguistics, 2019, http://dx.doi.org/10.18653/v1/d19-1074.

[19] G. Chen, D. Ye, Z. Xing, J. Chen, E. Cambria, Ensemble application of convolutional and recurrent neural networks for multi-label text categorization, in: 2017 International Joint Conference on Neural Networks, IJCNN, IEEE, 2017, pp. 2377–2383.

[20] Y. Li, Q. Pan, S. Wang, T. Yang, E. Cambria, A generative model for category text generation, Inform. Sci. 450 (2018) 301–315.

[21] I. Chaturvedi, Y.-S. Ong, I.W. Tsang, R.E. Welsch, E. Cambria, Learning word dependencies in text by means of a deep recurrent belief network, Knowl.-Based Syst. 108 (2016) 144–154.

[22] A.B. Dieng, F.J.R. Ruiz, D.M. Blei, Topic modeling in embedding spaces, Trans. Assoc. Comput. Linguist. 8 (2020) 439–453, http://dx.doi.org/10.1162/tacl_a_00325.

[23] M. Christian, M. Grottke, A systematic literature review of mining weak signals and trends for corporate foresight, J. Bus. Econ. 88 (5) (2018) 643–687.

[24] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (null) (2003) 993–1022.

[25] L. Pépin, P. Kuntz, J. Blanchard, F. Guillet, P. Suignard, Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets, Comput. Ind. Eng. 112 (2017) 450–458, http://dx.doi.org/10.1016/j.cie.2017.01.025.

[26] T. Gutsche, Automatic Weak Signal Detection and Forecasting (Master's thesis), University of Twente, Faculty of Behavioral Management and Social Sciences, 2018.

[27] H.I. Ansoff, Managing strategic surprise by response to weak signals, Calif. Manage. Rev. 18 (2) (1975) 21–33, http://dx.doi.org/10.2307/41164635.

[28] M. Holopainen, T. Marja, Weak signals: Ansoff today, Futures 44 (2012) 3, 198–205.

[29] N. Tabatabaei, Detecting Weak Signals by Internet-Based Environmental Scanning (Master's thesis), University of Waterloo, 2011.

[30] S. Kim, Y.-E. Kim, K.-J. Bae, S.-B. Choi, J.-K. Park, Y.-D. Koo, Y.-W. Park, H.-K. Choi, H.-M. Kang, S.-W. Hong, N.E.S.T.: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network futures, Futures 52 (2013) 59–73, http://dx.doi.org/10.1016/j.futures.2013.08.004.

[31] Y.J. Hong, D. Shin, J.H. Kim, High/low reputation companies' dialogic communication activities and semantic networks on Facebook: A comparative study, Technol. Forecast. Soc. Change 110 (2016) 78–92, http://dx.doi.org/10.1016/j.techfore.2016.05.003.

[32] D. Thorleuchter, D. Van den Poel, Technology classification with latent semantic indexing, Expert Syst. Appl. 40 (5) (2013) 1786–1795, http://dx.doi.org/10.1016/j.eswa.2012.09.023.

[33] D. Thorleuchter, D. Van den Poel, Protecting research and technology from espionage, Expert Syst. Appl. 40 (9) (2013) 3432–3440, http://dx.doi.org/10.1016/j.eswa.2012.12.051.

[34] D. Thorleuchter, D. Van den Poel, Weak signal identification with semantic web mining, Expert Syst. Appl. 40 (12) (2013) 4978–4985, http://dx.doi.org/10.1016/j.eswa.2013.03.002.
[35] M. Smith, Catalyzing social media scholarship with open tools and data, J. Contemp. East. Asia 14 (2) (2015) 87–96.
[36] S.-H. Yoo, H.-W. Park, K.-H. Kim, A study on exploring weak signals of technology innovation using informetrics, J. Technol. Innov. 17 (2) (2009) 109–130.
[37] R. Bamler, S. Mandt, Dynamic word embeddings, in: ICML, 2017.
[38] Z. Yao, Y. Sun, W. Ding, N. Rao, H. Xiong, Dynamic word embeddings for evolving semantic discovery, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 673–681, http://dx.doi.org/10.1145/3159652.3159703.
[39] A.B. Dieng, F.J.R. Ruiz, D. Blei, The dynamic embedded topic model, 2019, ArXiv abs/1907.05545.
[40] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 113–120, http://dx.doi.org/10.1145/1143844.1143859.
[41] E. Ekinci, S. c İlhan Omurca, Concept-LDA: Incorporating Babelfy into LDA for aspect extraction, J. Inf. Sci. 46 (3) (2020) 406–418, http://dx.doi.org/10.1177/0165551519845854.
[42] V. Rus, N. Niraula, R. Banjade, Similarity measures based on latent Dirichlet allocation, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 459–470.
[43] S. Poria, I. Chaturvedi, E. Cambria, F. Bisio, Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis, in: 2016 International Joint Conference on Neural Networks, IJCNN, 2016, pp. 4465–4473.
[44] J. Maitre, M. Ménard, G. Chiron, A. Bouju, N. Sidère, A meaningful information extraction system for interactive analysis of documents, in: 2019 International Conference on Document Analysis and Recognition, ICDAR, 2019, pp. 92–99.
[45] L. Kölbl, M. Grottke, Obtaining more specific topics and detecting weak signals by topic word selection, in: Reliability and Statistical Computing: Modeling, Methods and Applications, Springer International Publishing, Cham, 2020, pp. 193–206, http://dx.doi.org/10.1007/978-3-030-43412-0_12.
[46] J.C. Campbell, A. Hindle, E. Stroulia, Chapter 6 - Latent Dirichlet allocation: Extracting topics from software engineering data, in: C. Bird, T. Menzies, T. Zimmermann (Eds.), The Art and Science of Analyzing Software Data, Morgan Kaufmann, Boston, 2015, pp. 139–159, http://dx.doi.org/10.1016/B978-0-12-411519-4.00006-9.
[47] B. Lent, R. Agrawal, R. Srikant, Discovering trends in text databases, in: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997, pp. 227–230.
[48] S. Goorha, L. Ungar, Discovery of significant emerging trends, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 57–64, http://dx.doi.org/10.1145/1835804.1835815.
[49] H. Abe, S. Tsumoto, Trend detection from large text data, in: 2010 IEEE International Conference on Systems, Man and Cybernetics, 2010, pp. 310–315.
[50] X. Wang, P. Qiu, D. Zhu, L. Mitkova, M. Lei, A.L. Porter, Identification of technology development trends based on subject–action–object analysis: The case of dye-sensitized solar cells, Technol. Forecast. Soc. Change 98 (2015) 24–46, http://dx.doi.org/10.1016/j.techfore.2015.05.014.
[51] K. Nguyen, B.-J. Shin, S.J. Yoo, Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information, in: 2016 International Conference on Big Data and Smart Computing, BigComp, 2010, pp. 223–230, http://dx.doi.org/10.1109/BIGCOMP.2016.7425917.
[52] J.M. Gerken, M.G. Moehrle, A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis, Scientometrics 91 (3) (2012) 645–670, http://dx.doi.org/10.1007/s11192-012-0635-7.
[53] Y. Lee, S.Y. Kim, I. Song, Y. Park, J. Shin, Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis, Scientometrics 100 (1) (2014) 227–244, http://dx.doi.org/10.1007/s11192-013-1216-0.
[54] D. Thorleuchter, T. Scheja, D. Van den Poel, Semantic weak signal tracing, Expert Syst. Appl. 41 (11) (2014) 5009–5016, http://dx.doi.org/10.1016/j.eswa.2014.02.046.
[55] J. Chang, S. Gerrish, C. Wang, J.L. Boyd-graber, D.M. Blei, Reading tea leaves: How humans interpret topic models, in: Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, Curran Associates, Inc., 2009, pp. 288–296.
[56] D. Newman, J.H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, Association for Computational Linguistics, USA, 2010, pp. 100–108.
[57] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 262–272.
[58] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 399–408, http://dx.doi.org/10.1145/2684822.2685324.
[59] S. Yokoyama, H. Sanada, Logistic regression model for predicting language change, in: R. Kohler (Ed.), Issues in Quantitative Linguistics, 2009, pp. 176–192.
[60] D. Thorleuchter, D.V.D. Poel, Weak Signal Identification with Semantic Web Mining, Belgium 13/860, Ghent University, Faculty of Economics and Business Administration, 2013, Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium.
[61] J. Chuang, C.D. Manning, J. Heer, Termite: Visualization techniques for assessing textual topic models, in: Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 74–77, http://dx.doi.org/10.1145/2254556.2254572.
[62] C. Sievert, K. Shirley, LDAvis: A method for visualizing and interpreting topics, in: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 63–70, http://dx.doi.org/10.3115/v1/W14-3110.
[63] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings, 2013.
[64] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C.J.C. Burges, L. Bottou, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5–8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 3111–3119.
[65] Z. Lian, Exploration of the working principle and application of Word2Vec, Sci-Tech Inf. Dev. Econ. 2 (2015) 145–148.
[66] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 746–751.
[67] K. Hu, Q. Luo, K. Qi, S. Yang, J. Mao, X. Fu, J. Zheng, H. Wu, Y. Guo, Q. Zhu, Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis, Inf. Process. Manage. 56 (4) (2019) 1185–1203, http://dx.doi.org/10.1016/j.ipm.2019.02.014.
[68] A. Khatua, A. Khatua, E. Cambria, A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks, Inf. Process. Manage. 56 (1) (2019) 247–257, http://dx.doi.org/10.1016/j.ipm.2018.10.010.
[69] R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.
[70] S. Syed, M. Spruit, Selecting priors for latent Dirichlet allocation, in: 2018 IEEE 12th International Conference on Semantic Computing, ICSC, IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 194–202, http://dx.doi.org/10.1109/ICSC.2018.00035.
[71] H.M. Wallach, D.M. Mimno, A. McCallum, Rethinking LDA: Why priors matter, in: Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, Curran Associates, Inc., 2009, pp. 1973–1981.
[72] S. Syed, M. Spruit, Exploring symmetrical and asymmetrical Dirichlet priors for latent Dirichlet allocation, Int. J. Semant. Comput. 12 (2018) 399–423.
[73] D. Steel, C. McLaren, Chapter 33 - Design and analysis of surveys repeated over time, in: C. Rao (Ed.), Handbook of Statistics, in: Handbook of Statistics, vol. 29, Elsevier, 2009, pp. 289–313, http://dx.doi.org/10.1016/S0169-7161(09)00233-8.
[74] R. Frankland, A.D. Smith, J. Sharpe, R. Bhatia, S. Jarvis, P. Jakhria, G. Mehta, Calibration of VaR models with overlapping data, Br. Actuar. J. 24 (2019) e23, http://dx.doi.org/10.1017/S1357321719000151.