Journal of Classification
https://doi.org/10.1007/s00357-024-09476-0

Automatic Topic Title Assignment with Word Embedding

Gianpaolo Zammarchi[1] · Maurizio Romano[1] · Claudio Conversano[1]

Accepted: 29 May 2024

Abstract
In this paper, we propose TAWE (title assignment with word embedding), a new method to automatically assign titles to topics inferred from sets of documents. This method combines the results obtained from the topic modeling performed with, e.g., latent Dirichlet allocation (LDA) or other suitable methods and the word embedding representation of words in a vector space. This representation preserves the meaning of the words while allowing to find the most suitable word that represents the topic. The procedure is twofold: first, a cleaned text is used to build the LDA model to infer a desirable number of latent topics; second, a reasonable number of words and their weights are extracted from each topic and represented in n-dimensional space using word embedding. Based on the selected weighted words, a centroid is computed, and the closest word is chosen as the title of the topic. To test the method, we used a collection of tweets about climate change downloaded from some of the main newspapers accounts on Twitter. Results showed that TAWE is a suitable method for automatically assigning a topic title.

Keywords Automatic title assignment · Word embedding · Topic modeling · LDA · Centroid · Climate change · Twitter

1 Introduction

The amount of textual data produced every day by social media and other sources is enormous. Social media have become a trusted and popular information source for individuals all around the world. Many conventional newspapers post their main news on Twitter, one of the most used microblogging platforms, on a regular basis. Nowadays, people frequently use social media for either personal or professional activities, but the sheer number of sources of information might make it difficult to define the main topics that are covered. Therefore, there is a demand for effective techniques and tools that automatically summarize and evaluate information in online social networks. The study of natural language processing (NLP) enables machines to comprehend, evaluate, and analyze the meaning of human speech by fusing computational linguistics and artificial intelligence (Nadkarni et al., 2011).

Hereafter, we concentrate on probabilistic topic modeling (Blei, 2012), which can be described as a technique to identify word clusters (topics) in a collection of texts. Topic

✉ Gianpaolo Zammarchi
gianpaolo.zammarchi@unica.it

[1] Department of Economics and Business Science, University of Cagliari, Cagliari, Italy

Published online: 01 July 2024

models have been used for social media content analysis in a variety of tasks since the rise of social media in recent years, such as event monitoring (Vavliakis et al., 2013; Lau et al., 2012; Jeong et al., 2019), or content recommendations (Chen et al., 2010). Once the topics are extracted, it is necessary to assign a title to the set of words that compose the topic. In many cases, researchers manually choose the title to assign to the topic. In other cases, the assignment is automatic (Lau et al., 2011; Kozono & Saga, 2020; Truică & Apostol, 2021) and is based on word embedding (Bhatia et al., 2016). Word embedding represents words as vectors in an n-dimensional space and allows to perform several mathematical operations that are impossible to perform on text.

We propose a new methodology to automatically assign titles to topics using word embedding and apply it to text corpora concerning climate change. The main characteristic of our method called title assignment with word embedding (TAWE) is the definition of a centroid for each topic based on a list of words extracted from the topic and weighted according to their relevance. This method allows to find the most suitable representative among all the words that are included in a topic. Furthermore, it allows us to measure distance, hence similarity, of topics in order to directly compare both the topics and their sources (e.g., newspapers) numerically and/or visually. For example, using principal component analysis (PCA), information about topics and newspapers can be represented in $\mathcal{R}^2$ or $\mathcal{R}^3$ providing a useful tool for the interpretability of results. We apply TAWE to assign a title to topics obtained from tweets concerning climate change.

Climate change can be considered one of the greatest challenges of our time since the long-term consequences of an uncontrolled rise in temperatures across the planet would be dreadful. Some of the effects that we are already experiencing are glaciers melting, warming of seas and oceans, rising sea levels, and sudden and violent changes in weather or rainfall patterns. These events have important repercussions on our health, economy, and society in general. These effects are anticipated to worsen over the next years if substantial measures to reduce or stop human-induced greenhouse gas emissions are not taken soon. Nonetheless, even if the situation is very serious, people still struggle to see climate change as a problem (Van Lange & Huckelba, 2021). Among other things, the level of information plays a fundamental role in raising people's level of awareness about a topic such as healthy food (Wakefield et al., 2003) and smoking (Dumanovsky et al., 2010). Media are an important source of information, have a great influence on public opinion, and can help to spread knowledge about climate change. Moreover, the media may convey the initiatives that communities, governments, and people can take, as well as the repercussions of climate change. Although it is possible that worldwide climate strike movements, for example, contributed to an increase in media coverage of climate change in recent years, there may also have been a decline in coverage because other important issues emerged, such as the coronavirus disease 2019 (COVID-19) pandemic.

In the above-described framework, the main contributions of this paper are summarized as follows:

1) To introduce TAWE as a new method to automatically assign titles to topics extracted from a set of documents. This method combines the results obtained from the topic modeling and the word embedding representation of words in a vector space.

2) To introduce the topic modeling analysis of tweets on climate change, we assess trends in media coverage of climate change using tweets posted in English by main international newspapers from 2012 to 2021. We retrieved all tweets posted by three main newspapers based in the United Kingdom (UK) and three based in the United States (US). We show that, for the majority of newspapers, the number of tweets on climate change increased from

2014 to 2019, saw a sharp decrease in 2020, in correspondence with the emergence of the COVID-19 pandemic, and a subsequent rise in 2021.

3) To identify the main topics discussed in these tweets using topic modeling. To this aim, we applied latent Dirichlet allocation (LDA) and structural topic models (STM), which are widely used methods that allow inferring the structure of the hidden variables, i.e., the topics, given the documents. We show specific topics for each newspaper as well as common topics covered by different sources, such as politics, extreme weather, scientific reports, and gas emissions.

4) To present an in-depth analysis of common topics between newspapers.

The rest of the paper is organized as follows. Section 2 describes previous studies performing analyses on tweets related to climate change. Section 3 presents TAWE. In Sect. 4, we present the application of TAWE to the climate change data set and discuss results, while in Sect. 5 we present future steps and conclusions.

2 Background

A number of studies have used topic modeling on tweets related to climate change, using either tweets posted by different users or by specific accounts.

Dahal et al. (2019) analyzed tweets posted by different users and collected them using global warming-related keywords. LDA was utilized to identify the most popular topics and tweets posted by users living in different countries (Great Britain, Canada, Australia, and the US) and were compared to show that in the US there is less discussion on topics related to policies and useful actions to counteract climate change compared with other countries.

Sanford et al. (2021) examined more than 6000 English-language tweets to highlight the presence of topics. Using a Python library for topic modeling named CorEx (Correlation Explanation, Gallagher et al. 2017), they found that, after the publication of the IPCC report, one of the most discussed topics was food-related, i.e., meat consumption and dietary choices.

Calleo and Pilla (2022) collected a data set of more than 120,000 tweets spanning from 2007 to 2021. They used geographical information to select tweets posted only by Italian users and applied LDA to create 10 topics. Combining topics with geolocation data, they were able to assess the spatial distribution of topics over regions, stating that Topic 1 (climate change effects) was mainly discussed in the Italian regions of Lazio and Lombardy.

Effrosynidis et al. (2022) created a data set of tweets related to climate change merging three publicly available data sets (Credibility of Climate Change Denial in Social Media Data, Climate Change Tweets IDs Data, and Twitter Archive Data) with more than 15 million tweets spanning over 13 years (2006–2019). The data set was anonymized, and for more than a third of tweets, it was possible to add geolocation data. The authors conducted topic modeling on this data set showing that gas emissions, extreme weather, politics, and human intervention were among the main topics.

Ohtani (2022) focused on the importance of the aspect of biodiversity and used tweets to identify the most important words associated with this term over the last decade. The author also performed sentiment analysis using the NRC R package (Mohammad & Turney, 2013) to extract the main emotion expressed in a topic, discovering that the number of tweets followed an upward trend after 2018 and that their content can be classified as optimistic rather than pessimistic.

Other studies have investigated the role of bots (automated users) in spreading misinformation and denying the effects of climate change or have focused on tweets posted by specific accounts. Marlow et al. (2021) examined nearly seven million tweets posted in the time period following the US withdrawal from the Paris Agreement in June 2017, reporting that bot activity was higher in topics related to denial of climate change and its importance. Felaco et al. (2020) proposed a semi-automatic labeling of topics extracted from the FridaysForFuture official Twitter account. They resolved to use LDA in combination with social network analysis to define the content of each topic and their relationships. In particular, they showed that topics can be classified into three categories: climate change activists, strikes, and future dimensions. DePaula (2020) examined tweets posted by five US federal government agencies highlighting which were the 15 most important topics for each agency. The author also proposed a score for each topic ranging from 1 to 3, where 1 denotes a coherent topic judged by a human being, and 3 is a difficult-to-interpret topic. The study concluded that, despite the hostile political administration, agencies keep communicating about climate change, but some were influenced by the political administration in restricting climate change information.

To our knowledge, no study has applied topic modeling to tweets originating from a more reliable source of information compared with generic user accounts to evaluate how the interest of media towards climate change has evolved through the years. For such a reason, we focus our attention on tweets posted by traditional newspapers, which generally have a good reputation among people and whose sources of information are considered more reliable.

2.1 Topic Modeling

Topic modeling is a type of statistical model aimed at discovering the presence of “topics” in a (usually large) collection of documents (see Landauer and Dumais, 1997; Hofmann, 1999; Griffiths and Steyvers, 2002; Blei et al., 2003). The definition of a document is somehow broad, especially in NLP, where each written text (or a part of it) can be considered as a document. Moreover, the type and length of a document can vary significantly from one source to another, and therefore it is possible to use the word document for texts of a very different nature and size. When we think about the sheer number of comments, reviews, posts, tweets, etc., that are written every day by millions of users of many online platforms (social media, video, and photo sharing platforms), it is easy to understand that analyzing this type of data would be impossible without a technique that automates part of the work. Topic modeling helps to discover the latent semantic structures of a collection of documents (usually also referred to as corpora) and to have, at the end of the evaluation, a more precise idea of which topics are covered by the considered texts. In the last few years, topic modeling has become more and more important, and many scholars have developed new techniques to tackle the problem of identifying topics within a collection of texts (see, for example, Liu et al., 2016; Jelodar et al., 2019). Topic modeling can be applied in several fields, such as social network opinion analysis, health, education, marketing, and so on. Some of the most important techniques developed in the early days are the latent semantic analysis (LSA; Deerwester et al., 1990), the probabilistic LSA (PLSA), mainly based on word co-occurrence (Hofmann, 2001), and LDA, which is still one of the most well-known and used techniques today (Blei et al., 2003). In LDA the word latent refers to the topics, which are considered hidden variables, while the observed variables are the documents and the words composing those documents. Moreover, the Dirichlet allocation specifies which probability distribution we should use when we take into account the probabilistic part of the model, i.e., the Dirichlet distribution. Nonetheless, other techniques not strictly related to topic modeling can be used, such as non-negative matrix factorization (NNMF) (Lee & Seung, 2000). In nearly two decades, many new techniques have been developed starting from the LDA, such as the pachinko allocation model (PAM), which allows taking into account the correlation between topics (Li & McCallum, 2006). Another useful and promising approach is called stochastic block model (SBM). Basically, this method aims at generating nodes connected by edges where more densely connected parts can be identified as communities. It is an innovative method that uses network-based techniques to improve the quality of fit and the interpretability of topic modeling. This method is based on the generation of graphs (where words represent nodes and their relationship is expressed by edges) and can therefore be applied to topic creation. It was originally created by Holland et al. (1983) and recently adapted to be used for topic modeling (Gerlach et al., 2018). Moreover, since the model formulation is nonparametric, the model can be used without having to specify the number of topics a priori and, unlike LDA, which is based on a nonhierarchical clustering of the words alone, the model is based on a hierarchical clustering of both words and documents. For a more comprehensive review of literature on SBM for topic modeling, see, e.g., Lee and Wilkinson (2019).

There are several ways to classify topic modeling approaches, such as supervised, semi-supervised and unsupervised (depending on the type of data being used). Another important distinction can be found between probabilistic (e.g., LDA) and non-probabilistic methods (LSA, NNMF). In probabilistic topic modeling, a document is considered a mixture of topics, with an underlying generative process based on probability distributions. To create a new document, first, we need to pick a probability distribution over topics, which is the same for all documents in the collection, and then choose a word from the topic selected in the previous step. Every new document will pass through the same phases to create different documents according to the probability distribution of topics and words. The fact that the inferred hidden structure mirrors the theme organization of the collection is what makes topic models so useful. Each document in the corpus can be labeled and these labels can be used to perform explorative data analysis, information retrieval, and classification. Due to its widespread use and extensive testing in the last two decades, we decided to exploit LDA for our analysis.

2.2 Word Embedding

Word embedding is an approach aimed at converting words into numerical vectors (Bengio et al., 2000). In essence, it involves mathematically embedding a multidimensional space into a space with a significantly reduced number of dimensions. Basically, we are able to map each word in a vocabulary to a point in a vector space, using the numerical word representation. This is because the underlying hypothesis is that words that appear in the same context will have a similar meaning or will be somehow related. When we represent these words in a vector space, we expect to see that these related words will appear closer than unrelated words. In the last decade, the method has been increasingly adopted in several studies, after improvements in the vector representation and model's training speed, as well as hardware advancements that allowed for the profitable exploration of a larger parameter space. Word2vec, a word embedding toolkit developed in 2013 by Mikolov et al. (2013), allows users to train vector space models more quickly compared with earlier methods. The word2vec method has been widely applied and played a key role in generating interest in word embedding, from specific fields to broader experimentation, ultimately opening the door for its practical implementation. One disadvantage of word embedding is the collection of words with different meanings into a single representation. In other words, polysemy is not adequately managed. In the last few years, researchers have developed embeddings that take advantage of a word's context to discern between its alternative meanings in order to solve this issue (Devlin et al., 2018). Other problems are the possible presence of bias due to the text used to train these models (e.g., gender bias) and the fact that the models operate as a black box, i.e., it is not possible to understand what happened during the evaluation.

3 Title Assignment with Word Embedding

In this section, we describe the proposed title assignment with word embedding (TAWE) methodology to automatically assign titles to topics. We show how to extract K topics using LDA and how to transform words into vectors using word embedding. We decided to use word embedding due to its ability to preserve the semantic and syntactic relationships between words as well as the possibility to train it on large external corpora. In addition, while we could consider the creation of a simpler document-term matrix, this requires the availability of the original data set of documents. The choice to use word embedding in the TAWE method allows to apply it also in case only a list of words assigned to each topic, but not the original data set of documents, is available. By selecting the most representative words for each topic, we are able to summarize and compare the data collected and the information produced by LDA or any other topic modeling technique. Preliminarily, the original data set is subjected to a data-cleaning process composed of different preprocessing steps. Those steps include the conversion of each word to lowercase letters, removal of punctuation, numbers, URLs, symbols, and stop words, and removal of those words that are present in the large majority of the texts (e.g., keywords used to select a set of documents).

3.1 Topics Extraction with LDA

Let $D = \{d_1, d_2, \ldots, d_D\}$ be the collection of documents used for the analysis. Each d_i is considered as a set of words of different length, that is $d_i = \{w_1, w_2, \ldots, w_N\}$ with $i = \{1, \ldots, D\}$.

The main goal when we use LDA, and in general topic modeling, is to infer the structure of the hidden variables given the documents. Statistically speaking, the problem is to compute the posterior distribution of the hidden variables given the documents

$$p(\varphi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(\varphi_k) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\varphi_{1:K}, z_{d,n}) \right)$$

where K represents the number of topics, D is the number of documents, and N is the number of words. Moreover, φ_k is the per-corpus topic distribution, θ_d is the per-document topic proportions, $z_{d,n}$ is the per-word topic assignment and $w_{d,n}$ is the observed word (Blei, 2012).

When LDA is applied, two matrices are produced: the word-topic matrix $\mathbf{T}^{W}_{(K,N)}$, and the document-topic matrix $\mathbf{T}^{D}_{(K,D)}$. The distribution for the two matrices is the Dirichlet distribution, but with two different hyperparameters (α and β). In addition to α and β, the number of topics K is another user-defined parameter of the model. More formally, LDA assumes the following generative process

Step 1
$\theta_d \sim \mathrm{Dir}(\alpha)$, with $d \in \{1, \ldots, D\}$

Step 2
$\varphi_k \sim \mathrm{Dir}(\beta)$, with $k \in \{1, \ldots, K\}$

Step 3

For each new word $w_{d,n}$ to be generated:

(a) Choose a topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$

(b) Choose a word $w_{d,n} \sim \text{Multinomial}(\varphi_{z_{d,n}})$.

The hyperparameter α affects the number of topics in a document since if we choose a low α, the document will only consist of a few topics (and vice versa). The same argument applies to β to identify the most probable terms for a topic (Jacobi et al., 2016).

The outputs of LDA, i.e., the sets of words characterizing each topic and the proportion of times each word appears in the topic, are the input of the TAWE algorithm. The latter is composed of two steps: (1) Topic labels' assignment (Sect. 3.2, Algorithm 1), and (2) Topics' labels refinement (Sect. 3.3, Algorithm 2).

3.2 Topics Labels' Assignment

After applying LDA, words assigned to each topic have to be converted into vectors. LDA also provides probabilities for each word to be extracted from a topic. These probabilities reflect the importance of each word in that topic and are here used as weights. We hereby consider word embedding as an alternative method to find vectors and introduce a new method to assign a title to a topic using word embedding representation. Since each word can be expressed in vector representation, operations such as addition, subtraction, and others can be performed between words. These properties allow us to compute the centroid for a group of words. The procedure is straightforward. First, we define W as the set of unique words of all the M words in the K topics, where $M < N$ (N is the total number of words in all D documents).

$$W = \bigcup_{k=1}^{K} \bigcup_{m=1}^{M} \{w_{k,m}\} \tag{1}$$

The procedure can be carried out for any arbitrary $M \in [1; N]$, but for M = 1, even though the procedure would still be feasible, it would present obvious results. Next, we focus on a word embedding distributed representation of word meanings

$$\boldsymbol{E}_{d_{\text{emb}} \times d_{\text{corpus}}} \in \mathbb{R}^{d_{\text{emb}}} \tag{2}$$

s.t. $|W| - |\boldsymbol{E} \cap W| \simeq 0$, where d_{emb} is the number of dimensions of the vectors chosen for the word embedding, while d_{corpus} represents the number of words in the training corpus set of the distributed representation.

Both d_{emb} and M impact on the results of the procedure. First, increasing d_{emb} improves the word embedding representation, but at the cost of increasing the computational complexity in terms of memory requirements for processing (Sindhu & Seshadri, 2021). In view of that, following Yamada et al. (2020), a pre-trained embedding with $d_{\text{emb}} = 100$ has been carried out. Second, the number of words M in the K topics impacts on the computational complexity of the method as well. Thus, the heterogeneity of words within each topic is another issue to be considered. In Sect. 3.3, it is shown how to choose a reasonable number of words M in each topic that preserves the information content of the topic itself but allows the procedure to assign labels to topics that are as different as possible from each other.

For a topic t_i, where $i \in \{1, ..., K\}$, we consider the matrix $\boldsymbol{E}_{t_i} \subset \boldsymbol{E}$ composed by the set of the $\omega_{t_i} \in t_i$ words represented in d_{emb} dimensions. Each of these M words is associated with a corresponding probability to occur in a topic, obtained in the topic modeling procedure, and used in this step as a vector of unnormalized weights $\boldsymbol{\pi}_{t_i} = \{p_1, \ldots, p_m, \ldots, p_M\} \in t_i$ and $p_m \in [0; 1]$ in the word embedding procedure. Hence, while considering the set $\omega_{t_i} \in t_i$ as a cluster of embedding vectors, we first need to find a vector of d_{emb} dimensions that is as much as possible representative of the cluster (Xu & Tian, 2015). Jin and Han (2010) have proved that using the median allows to compute centroids that are more robust to noise and outliers. The first considers the "most central" element within a cluster as the center of gravity, in this way reducing the possible effect of outliers. In our framework, the centroid is computed as the median among the weighted embedded word vectors:

$$c_{t_i} = \text{median}(\boldsymbol{E}_{t_i} \cdot \boldsymbol{\pi}_{t_i}) \tag{3}$$

Next, the label l_{t_i} for a specific topic t_i is found by searching for the word w_{t_i} included in $\boldsymbol{E}_{t_i}$ for which the cosine similarity $S_{C_{t_i}}$ is maximized, namely

$$l_{t_i} = \arg\max_{w_{t_i}} S_{C_{t_i}}(c_{t_i}, w_{t_i}) = \arg\max_{w_{t_i}} \left(\frac{c_{t_i} \cdot w_{t_i}}{||c_{t_i}|| \; ||w_{t_i}||} \right) \tag{4}$$

with $w_{t_i} \in \boldsymbol{E}_{t_i}$. The cosine similarity $S_{C_{t_i}}$ measures how similar the vector w_{t_i} and the centroid c_{t_i} are. It equals one if the two vectors are exactly the same, while it approaches one as long as their similarity increases.[1] Although, in principle, $S_{C_{t_i}}$ might be negative ($-1 \leq S_{C_{t_i}} \leq 1$), in textual analysis the range of possible values of the cosine similarity is usually in the interval [0; 1].

The assignments in Eqs. 3 and 4 can be carried out for each topic t_i stemming out from the LDA model and can be formalized into an algorithm. Starting from the output of LDA (Sect. 3.1), recalling that ω_{t_i} is the list of words for each topic t_i and $\boldsymbol{\pi}_{t_i}$ is a vector of weights associated with each word, we define $\omega^*_{t_i}$ and $\pi^*_{t_i}$ as a subset of size M of the most frequent words and weights produced by the topic modeling analysis. Next, we considered $\vec{\omega}^*_{t_i}$ as the vector representation of the set of words $\omega^*_{t_i}$ and for each of the K topics t_i we computed the centroid c_{t_i} and then we looked for the vector $\vec{w}_{t_i} \in \vec{\omega}^*_{t_i}$ which is closest to the centroid c_{t_i}. By repeating this procedure for each topic t_i, we get (K centroids and) K labels included in the vector $\boldsymbol{l} = \{l_{t_1}, \ldots, l_{t_K}\}$. Notationally, the main steps of the TAWE algorithm are summarized in Algorithm 1.

Algorithm 1 TAWE titles assignments.

```
Input: $[\omega_{t_1}, \omega_{t_2}, \cdots, \omega_{t_K}]$, $[\boldsymbol{\pi}_{t_1}, \boldsymbol{\pi}_{t_2}, \cdots, \boldsymbol{\pi}_{t_K}]$        Output of LDA
for $i = 1$ to $K$ do
    $\omega^*_{t_i} = [w_1, w_2, \cdots, w_M]$
    $\pi^*_{t_i} = [p_1, p_2, \cdots, p_M]$
    for $m = 1$ to $M$ do
        $f : (w_m) \rightarrow \vec{w}_m, \ \vec{w}_m \in \mathbb{R}^{d_{\text{emb}}}$        (2)
    end for
    $\vec{\omega}^*_{t_i} = [\vec{w}_1, \vec{w}_2, \cdots, \vec{w}_M]$
    $c_{t_i} = \text{median}(\vec{\omega}^*_{t_i} \cdot \pi^*_{t_i})$        (3)
    $l_{t_i} = \arg\max_{\vec{w}_m} S_{C_{t_i}}(c_{t_i}, \vec{\omega}^*_{t_i})$        (4)
end for
$\boldsymbol{l} = [l_{t_1}, l_{t_2}, \cdots, l_{t_K}]$
Output: $\boldsymbol{l}$
```

[1] Conversely, the cosine similarity between two orthogonal vectors is zero.

Furthermore, considering that likewise a word, a label is composed of a numeric vector that can be represented in a multidimensional space, each topic can be visualized in a post hoc analysis for improving results interpretation.

3.3 Topics' Labels Refinement

Algorithm 1 returns a set of labels for each topic from 1 to K using a specific value of m. Let L_{MK} be the matrix containing the titles for all the K topics using all values of m in the range from 1 to M so that, e.g., $l_{\cdot 1}$, is the column vector of labels for Topic 1 using $m = \{1, \ldots, M\}$.

$$L_{MK} = \begin{bmatrix} l_{11} & l_{12} & \ldots & l_{1K} \\ l_{21} & l_{22} & \ldots & l_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ l_{M1} & l_{M2} & \ldots & l_{MK} \end{bmatrix}$$

After obtaining several possible titles (i.e., labels) for each topic, TAWE's title assignment can be done according to two alternative criteria.

The default criterion is based on the idea that the label characterizing a specific topic is the one that appears more frequently among the set of possible labels characterizing the topic itself. In view of that, it evaluates all the possible titles assigned to a topic ($l_{1K}, l_{2K}, \ldots, l_{MK}$, with a number of words m varying from 1 to M) and selects the title as the word that appears more frequently in all K titles using the "majority vote" criterion. For a set of labels/titles $l_{\cdot,k}$ characterizing a topic, the refined set $\boldsymbol{l}_{\text{ref}}$ is obtained as follows:

$$\boldsymbol{l}_{\text{ref}} = \bigcup_{k=1}^{K} \{\psi(l_{\cdot k})\}$$

where $\psi(\cdot)$ is the mode function. If multiple modes are found within a specific topic, following a parsimonious approach the refined label is that corresponding to the vector $l_{\cdot,k}$ having the minimum length, i.e., minimum m.

The second criterion considers the differences among the labels assigned to the different topics. The goal is obtaining topics that are as much as possible different in content one from another. Let $\boldsymbol{l}_m$ be the set of assigned labels to the K topics for a specific number of words $m \in \{1, 2, \ldots, M\}$, which is contained in $\boldsymbol{E}$, and therefore can be represented by numerical vectors. The matrix $\boldsymbol{E}_{\boldsymbol{l}_m} \subset \boldsymbol{E}$ is composed by the set of all the labels $\boldsymbol{l}_m$ represented in d_{emb} dimensions. The square matrix $\boldsymbol{S}$ of dimensions $|\boldsymbol{l}_m| \times |\boldsymbol{l}_m|$, and the vector $\boldsymbol{S}'_m$ of elements of the lower triangular part of $\boldsymbol{S}$ s.t.

$$\boldsymbol{S} = \left(\begin{array}{ccccc} & l_{m,1}, & l_{m,2}, & \cdots & l_{m,K} \\ l_{m,1} & 1 & \cos(l_{m,1}, l_{m,2}) & \cdots & \cos(l_{m,1}, l_{m,K}) \\ l_{m,2} & \cos(l_{m,2}, l_{m,1}) & 1 & \cdots & \cos(l_{m,2}, l_{m,K}) \\ \cdots & \cdots & \cdots & 1 & \cdots \\ l_{m,K} & \cos(l_{m,K}, l_{m,1}) & \cos(l_{m,K}, l_{m,2}) & \cdots & 1 \end{array} \right)$$

with $l_{m,1}, \ldots, l_{m,K} \in \boldsymbol{E}_{l_m}$ and $\boldsymbol{S}'_m = S_{ij} \; \forall i \geq j$. Recalling that TAWE (Alg. 1) uses a subset $\omega^*_{t_i}$ of size m of the most frequent words produced by the topic modeling analysis, we consider a function $\phi(\boldsymbol{S}'_m)$ that returns a measure of variability (for example, the standard deviation) of all the elements of $\boldsymbol{S}'_m$. While running the algorithm with any subset of size m, it is possible to produce different $\boldsymbol{S}'_m$ and, hence, select m, and the corresponding vector of labels $\boldsymbol{l}_{\text{ref}}$, that allows from maximum variability among the contents of the topics. In this case, the refined set of labels concerning topics characterized by the m most important words is defined as follows:

$$\boldsymbol{l}_{\text{ref}} = \arg\max_m \phi(\boldsymbol{S}'_m)$$

As a result, $\boldsymbol{l}_{\text{ref}}$ is the best labeling set for those topics.

Algorithm 2 explains in detail the necessary steps to carry out the refinement phase. In both cases (“majority” or “variability” criterion), the $M \times K$ matrix of possible topics’ titles reduces to a vector of length K of refined titles.

Algorithm 2 TAWE’s label refinement.

```
Input: $\boldsymbol{L}_{\boldsymbol{MK}}$ & criterion = (“majority” OR “variability”)

if criterion = “majority”
    $\boldsymbol{l}_{\text{ref}} = \bigcup_{k=1}^{K} \{\psi(l_{\cdot k})\}$
else
    for $m = 1$ to $M$ do
        Compute $\boldsymbol{S}'_m = S_{ij}$        $\forall i \geq j$

    $\boldsymbol{l}_{\text{ref}} = \arg\max_m \phi(\boldsymbol{S}'_m)$

Output: $\boldsymbol{l}_{\text{ref}}$
```

4 Application to the Climate Change Data Set

In this section, we present the results of the analysis performed using TAWE on a climate change data set. Before starting with TAWE, we present a descriptive analysis to show the composition of the topics and, in particular, the possible overlap and distance between them as well as the most important words for each topic.

4.1 Data Processing

Before starting with the topic extraction, data has to be downloaded, filtered, and cleaned. Hereafter, we present these three preliminary steps along with some additional exploratory analyses.

Data was collected by downloading all the 3,275,499 tweets posted from January 2012 to December 2021 on the official Twitter account of three widely known newspapers from the UK (The Guardian, The Independent, and The Mirror) and three from the US (The New York Times, The Washington Post and The Wall Street Journal). In order to select tweets related to climate change and environmental awareness, only tweets reporting at least one of the following keywords related to climate change: climate change, sustainability, earth day, plastic free, global warming, pollution, environmentally friendly, or renewable energy, were retained, leading to a total of 11,155 tweets.

In the first part of the preliminary analysis, we compared the evolution of the number of climate change tweets, as well as the number of total tweets posted by the six newspapers in the period of collection (from 2012 to 2021). In Fig. 1, the plot shows the number of tweets on climate change posted each year from January 1, 2012, to December 31, 2021, by the six newspapers (solid lines) and the overall number of tweets posted during the same period (dashed lines).

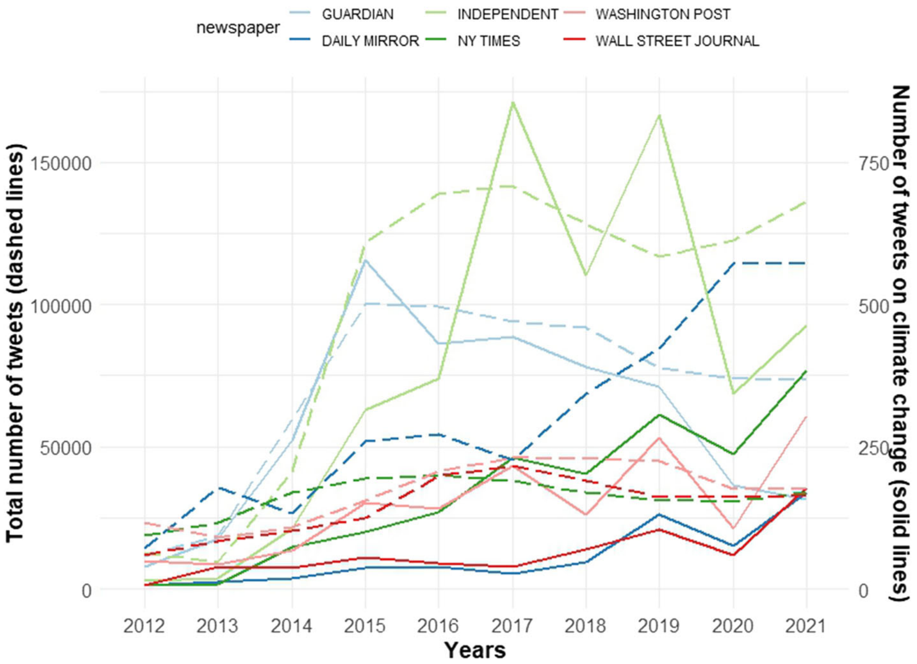

Fig. 1 Overall number of tweets (dashed lines, left Y axis) and tweets on climate change (solid lines, right Y axis) posted by the six newspapers from 2012 to 2021

Most of the newspapers increased the number of tweets on climate change from 2014 to 2019, but in 2020, after the COVID-19 outbreak, this number dropped significantly because all attention was focused on the aftermath of the pandemic. In 2021, we can observe a start of recovery towards the pre-pandemic values, which will be reached again within a few years. It is also important to point out that, unlike other newspapers, The Guardian showed a downward trend (started in 2016) that was probably accelerated by the pandemic but saw the number of tweets on climate change dropping even further in 2021. Figure 2 shows the percentage of change for both the overall number of tweets and climate change tweets posted by each newspaper from 2012 to 2021. In this way it is possible to carry out an evaluation both for a single newspaper, noting, e.g., that the New York Times has increased the number of tweets related to climate change more than the overall number of tweets, or to carry out an evaluation from a different point of view to reveal that, e.g., in 2020 all newspapers have reduced the interest in climate change (probably due to the COVID-19 outbreak).

To gain insight into the relationship between the number of tweets posted in general by a newspaper and the number of tweets related to climate change, we analyzed the association between these values for each newspaper using Spearman's correlation analysis. The number of tweets on climate change was positively and significantly correlated with the overall number of tweets for four of the six newspapers (The Guardian, Spearman's rho = 0.95, p < 0.001; The Mirror, Spearman's rho = 0.95, p < 0.001; The Independent, Spearman's rho = 0.76, p = 0.016; The Washington Post, Spearman's rho = 0.70, p = 0.031) but not for The New York Times (Spearman's rho = 0.18, p = 0.63) or The Wall Street Journal (Spearman's rho = 0.49, p = 0.15) (Zammarchi et al., 2022).

A further analysis was aimed at identifying changepoints (unexpected changes in a time series) in the number of tweets posted by each newspaper. For this analysis, we used the

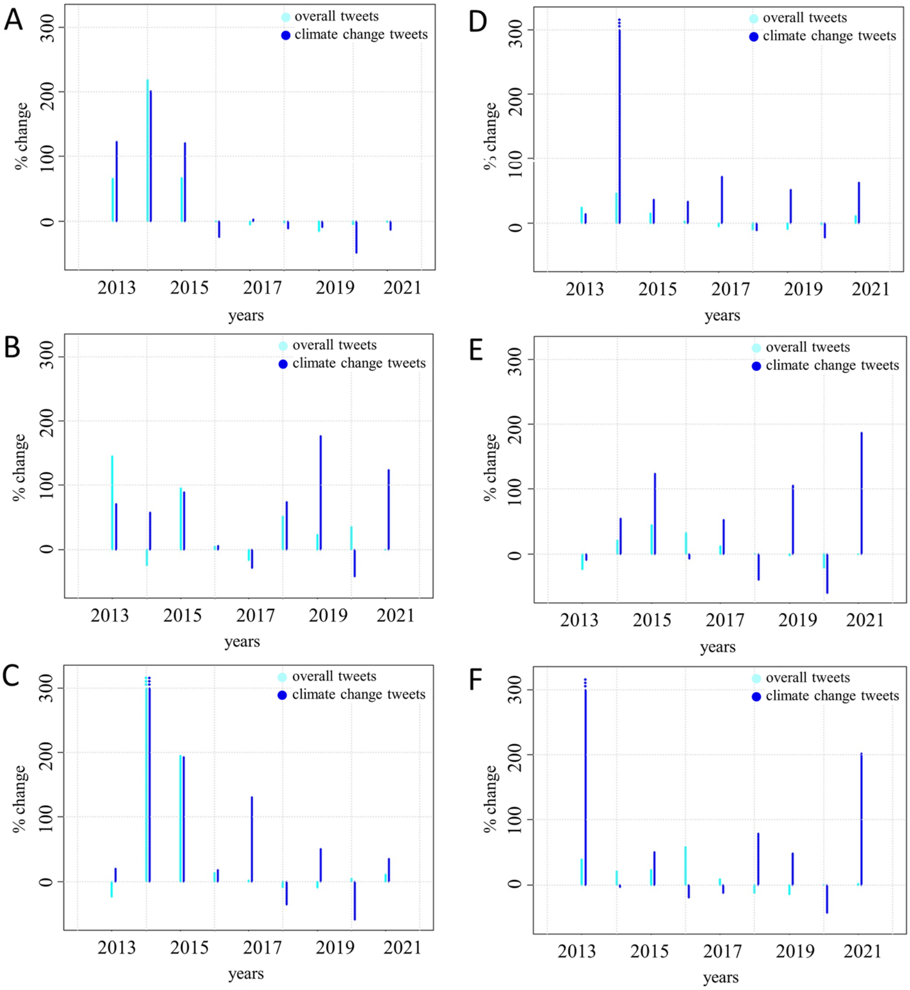

Fig. 2 Comparison of the percentage change in the number of tweets posted by the six newspapers from 2012 to 2021

changepoint package (Killick & Eckley, 2014) in R (R Core Team, 2020). We resolved to use this analysis because, in many cases, it is reasonable to think that one or more exogenous events occur, and so it will be possible to observe *b* breakpoints, each one corresponding to a shift in the mean value. The changepoint R package allows to use of different penalty criteria, such as the Bayesian information criterion (BIC) or the Akaike information criterion (AIC). We estimated the breakpoints on data aggregated by trimester, using the binary segmentation (BinSeg) method (Scott & Knott, 1974) implemented in the package. As shown in Fig. 3, we identified several structural breaks and different patterns for each newspaper.

Although there is a wide discrepancy in the number and position of the changepoints, it can be seen that at the end of 2018/beginning of 2019, the newspapers have some points in common. In particular, the three UK newspapers have the first quarter of 2019 as a common changepoint (Fig. 3, letters A, B, C), while the three US newspapers have the third quarter of 2018 as a common changepoint (Fig. 3, letters D, E, F). This changepoint corresponded to an

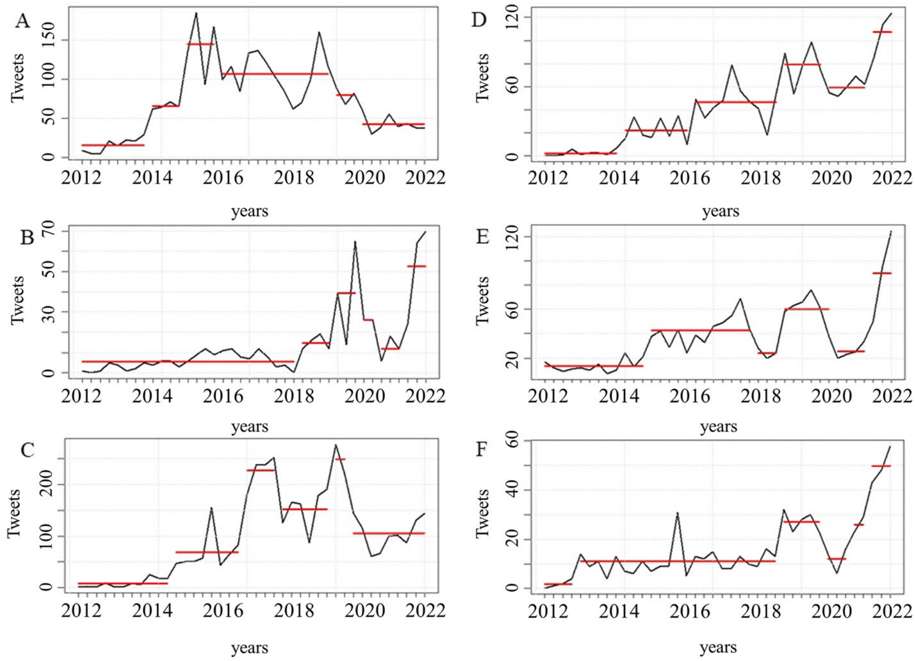

Fig. 3 Structural changes in the time series of tweets related to climate change. **A** The Guardian, **B** The Mirror, **C** The Independent, **D** The New York Times, **E** The Washington Post, **F** The Wall Street Journal. The red line represents the years between two breakpoints

increased number of climate change-related tweets for all newspapers except The Guardian. This finding can probably be explained by the media coverage of the strikes launched by the Fridays for Future movement, starting from the end of 2018. In 2020, during the Covid-19 outbreak, all newspapers observed one or more changepoints corresponding to a drastically reduced number of climate change tweets. Finally, in the first trimester of 2021, four out of six newspapers (Fig. 3, letters B, D, E, F) observed a changepoint corresponding to a substantial increase of climate tweets. Conversely, The Independent only observed a slight increase in tweets compared to the previous year, and The Guardian observed a steady trend.

4.2 Topics Extraction

In this section, we exploited the topic modeling approach to identify and analyze the main topics discussed by newspapers in their tweets. For each newspaper, we carried out some text preprocessing operations such as removal of URLs, numbers, stopwords, and keywords used to collect tweets. We build an LDA model for each newspaper on the cleaned text using the Gensim python library (Řehůřek & Sojka, 2011). In order to determine the appropriate number of topics, we tested a range of 2–20 topics and found the optimal number based on perplexity and coherence scores (Table 10). The perplexity is a metric that allows to evaluate the goodness-of-fit of an LDA model, with a lower perplexity score indicating better generalization performance (Blei et al., 2003). The coherence score reflects the semantic relatedness between words in a topic, with higher values indicating that a topic is internally

Table 1 Newspapers and topic coherence

Newspaper	Mean coherence (standard deviation)
The Guardian	0.61 (0.04)
The Daily Mirror	0.45 (0.08)
The Independent	0.58 (0.04)
The New York Times	0.44 (0.05)
The Washington Post	0.53 (0.05)
The Wall Street Journal	0.46 (0.02)

consistent (Mimno et al., 2011). For each newspaper, an optimal number of topics was chosen based on the evaluation of these two metrics by the authors using a criterion similar to the elbow method for clustering (Thorndike, 1953). We compared the plotted values for perplexity and coherence in order to find the number of topics that can be considered suitable for both methods. Finally, pyLDAvis (Sievert & Shirley, 2014; Mabey, 2021) was used to obtain a graphical representation of the topics in a bidimensional space. The mean topic coherence for topics identified for each newspaper is shown in Table 1.

Figures 4 and 5 show a bidimensional representation of the LDA models for UK- and US-based newspapers, respectively. Each newspaper has an optimized number of topics. In these plots, the topics are represented using circles, and the size of these circles is proportional to their prevalence in the corpus, while the distance is computed using the Jensen-Shannon divergence.

Tables 2, 3, and 4 show the top 10 words for topics identified in UK newspapers, while Tables 5, 6, and 7 for topics identified in US newspapers. As shown in Table 2, some of the topics identified in tweets from The Guardian are related to politics (Topic 2), energy (Topic 3), research studies (Topic 4), or consequences of climate change (Topic 5).

Table 3 shows topics identified in tweets from The Daily Mirror. In this case, we found most topics to be focused on threats related to the consequences of climate change (Topics 1, 4, 6, 8, 11, and 12). Other topics were related to activism (Topic 5) or politics (Topic 7). Table 4 shows topics identified in the last UK-based newspaper, The Independent. Topics identified in this newspaper show similarities with those identified in tweets posted by the

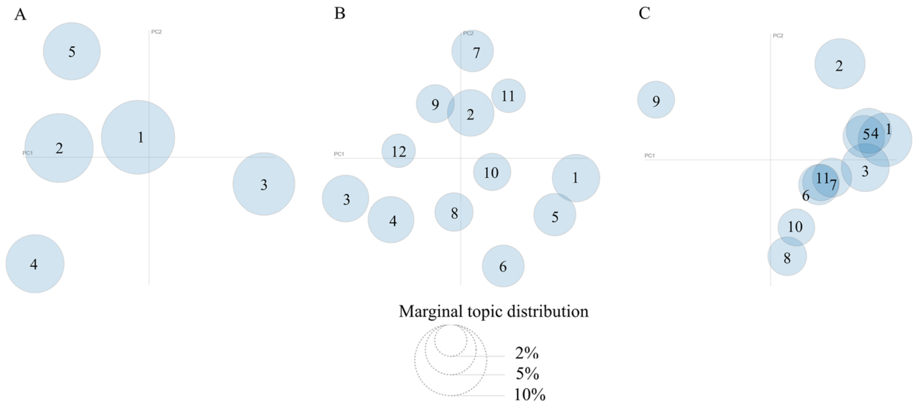

Fig. 4 Topic visualization in two dimensions for the UK newspapers (**A** The Guardian; **B** The Daily Mirror; **C** The Independent)

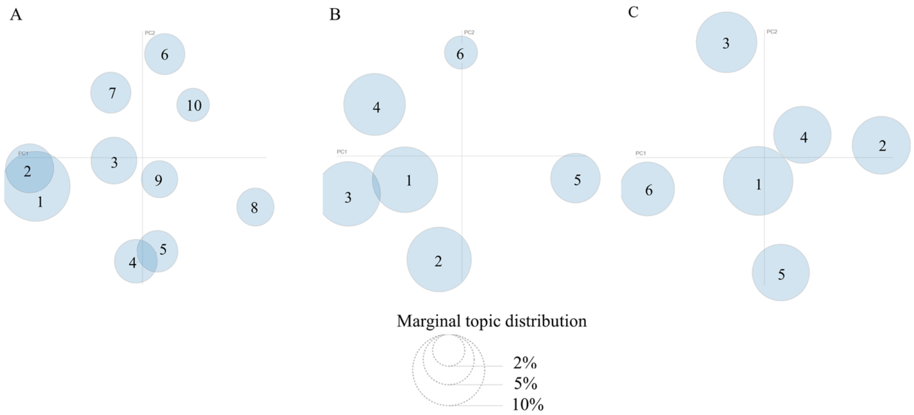

Fig. 5 Topic visualization in two dimensions for the US newspapers (**A** The New York Times; **B** The Washington Post; **C** The Wall Street Journal)

other two journals, such as politics (Topic 10), activism (Topic 2), energy (Topic 6), research studies (Topic 11), and consequences of climate change (Topics 1 and 8).

As regards US-based newspapers, Table 5 shows the top 10 words for topics identified in tweets posted by The New York Times. In this case, the majority of topics were related to politics (Topics 3, 5, and 7), consequences of climate change (Topics 1, 4, and 10), or both (Topics 2 and 8).

As shown in Table 6, topics identified in tweets posted by The Washington Post were related to politics (Topics 1 and 2) or the interaction between politics and either scientists (Topic 3). For this newspaper, we also identified topics related to actions to partake (Topics 5 and 6).

Finally, topics identified in tweets posted by The Wall Street Journal were mostly related to politics (Topics 1 and 4) or energy sources and emissions (Topics 2, 3, 5 and 6).

Overall, politics was one of the few themes covered by all newspapers and the one to which the majority of topics was related to. A lower number of topics was related to the effects of climate change as well as actions that government and people should partake to address this problem.

Table 2 Top ten words for topics identified in The Guardian tweets

Topic	Top ten words
Topic 1	view (0.018), editorial (0.014), crisis (0.013), people (0.010), delhi (0.009) video (0.009), paris (0.009), fight (0.008), don (0.008), key (0.008)
Topic 2	trump (0.036), london (0.026), court (0.013), plan (0.011), year (0.010) government (0.009), health (0.008), save (0.008), strikes (0.008), action (0.008)
Topic 3	energy (0.045), renewable (0.040), share (0.010), threat (0.010), years (0.009) australia (0.009), report (0.009), scientists (0.008), experiences (0.008), cars (0.007)
Topic 4	study (0.043), finds (0.028), linked (0.016), report (0.016), risk (0.014) school (0.014), work (0.009), fire (0.009), impact (0.009), tackling (0.009)
Topic 5	fight (0.017), warns (0.012), study (0.010), leaders (0.009), politicians (0.009) worse (0.009), worst (0.009), war (0.008), schools (0.008), dangerous (0.008)

Table 3 Top ten words for topics identified in The Daily Mirror tweets

Topic	Top ten words
Topic 1	experts (0.029), warn (0.027), years (0.026), david (0.020), sir (0.018) attenborough (0.018), floods (0.016), die (0.014), cop (0.014), levels (0.014)
Topic 2	johnson (0.023), years (0.022), act (0.018), polar (0.018), bears (0.018) cop (0.016), people (0.016), jailed (0.016), plastic (0.013), warns (0.013)
Topic 3	find (0.032), impact (0.031), live (0.028), century (0.028), terrible (0.014) killed (0.014), thousands (0.014), australia (0.014), camels (0.014), feral (0.014)
Topic 4	facing (0.022), cop (0.020), major (0.020), landmarks (0.020), temperatures (0.020) rise (0.018), countries (0.018), storms (0.018), wars (0.018), swell (0.018)
Topic 5	forever (0.018), years (0.018), protesters (0.015), chaos (0.015), die (0.015) risk (0.012), motorway (0.012), driver (0.012), busy (0.012), furious (0.012)
Topic 6	fight (0.023), sea (0.023), save (0.020), coastal (0.016), levels (0.016) arrested (0.016), threat (0.014), towns (0.014), vows (0.014), protesters (0.014)
Topic 7	day (0.019), extinction (0.019), heatwave (0.014), wont (0.014), images (0.011) celebs (0.011), rebellion (0.011), banner (0.011), put (0.011), cenotaph (0.011)
Topic 8	donald (0.020), fears (0.020), heat (0.014), trumps (0.014), strategy (0.014) turned (0.014), wildfires (0.014), face (0.014), trump (0.014), banned (0.014)
Topic 9	humans (0.024), polar (0.019), bear (0.019), weather (0.018), ways (0.018) shrinking (0.018), affects (0.018), wild (0.018), reindeer (0.012), adapt (0.012)
Topic 10	act (0.028), crisis (0.027), coronavirus (0.023), proves (0.021), year (0.021) due (0.019), bad (0.018), weather (0.016), death (0.016), extinction (0.013)
Topic 11	prince (0.037), due (0.022), charles (0.022), trump (0.021), school (0.019) greta (0.015), thunberg (0.015), fears (0.013), sea (0.013), warned (0.013)
Topic 12	brits (0.036), hotspots (0.027), favourite (0.022), due (0.022), ruined (0.018) holiday (0.018), birds (0.016), plastic (0.016), british (0.016), island (0.016)

4.3 Topics Auto-labeling

By applying Algorithm 1 (Sect. 3.2) to each newspaper, it was possible to automatically assign a label to each of the extracted topics. As described in Sect. 3.3, we obtained the title for each topic following the majority vote approach. After generating a set of titles for values of $m = 10$ up to $M = 100$ by steps of 10, we selected the most frequently assigned title for each topic. Alternatively, it is possible to calculate variance and median of all sets of labels produced while varying the size of the set of most frequent words m used to estimate the label of a single topic. Additionally, in Table 11 in the Appendix, we report an example of one newspaper while varying m. By plotting each title in bidimensional space (after applying to the word vectors some dimensionality reduction techniques such as PCA), we were also able to get a general overview of the results. Figure 6 shows whether or not some titles are shared by two or more newspapers (and which these newspapers are, e.g., three out of five titles of The Guardian are in common with The Independent) or, conversely, when titles are only attributed to one newspaper (when points are not overlapped with any other). Figure 7, on the other hand, shows the potential of TAWE, which allows to repeat the procedure in a hierarchical manner to obtain titles according to the required level of detail. For example, if we wish to obtain a single title for each newspaper, it is possible to join the titles provided for

Table 4 Top ten words for topics identified in The Independent tweets

Topic	Top ten words
Topic 1	scientists (0.046), made (0.034), europe (0.032), threat (0.027), earth (0.024) major (0.020), fossil (0.020), future (0.020), devastating (0.019), reveal (0.018)
Topic 2	report (0.046), extinction (0.031), action (0.031), government (0.026), free (0.022) years (0.019), day (0.017), set (0.017), companies (0.016), experts (0.016)
Topic 3	children (0.048), friendly (0.028), speech (0.027), environmentally (0.027), city (0.027) brexit (0.025), record (0.024), obama (0.023), forests (0.021), ocean (0.021)
Topic 4	opinion (0.087), plans (0.031), deniers (0.031), green (0.021), billion (0.020) launches (0.019), debate (0.017), epa (0.016), warns (0.015), bill (0.015)
Topic 5	fight (0.055), boris (0.044), johnson (0.042), china (0.039), time (0.032) weather (0.025), trees (0.022), research (0.022), plan (0.022), linked (0.022)
Topic 6	energy (0.078), renewable (0.069), due (0.023), warn (0.023), risk (0.020) sea (0.017), coal (0.017), power (0.017), britain (0.015), germany (0.015)
Topic 7	watch (0.039), live (0.031), house (0.027), levels (0.025), reveals (0.024) worlds (0.024), country (0.022), cities (0.022), emissions (0.022), white (0.021)
Topic 8	attenborough (0.026), scientists (0.025), activist (0.024), david (0.024), prince (0.023) drop (0.021), warning (0.021), save (0.018), york (0.017), isn (0.015)
Topic 9	plastic (0.062), people (0.057), paris (0.025), make (0.024), water (0.018) leaders (0.018), agreement (0.016), young (0.016), theresa (0.015), british (0.015)
Topic 10	trump (0.191), donald (0.064), carbon (0.024), trumps (0.024), fight (0.022) administration (0.020), environment (0.018), eu (0.017), top (0.015), fighting (0.015)
Topic 11	study (0.095), finds (0.047), crisis (0.039), suggests (0.031), year (0.023) life (0.023), summit (0.021), cop (0.018), effects (0.017), denier (0.017)

the various topics into a single, general title. In this way, thanks to the graphic representation as well, it would be possible to evaluate the difference between the macro-themes covered by the various newspapers according to the "distance" obtained in the two-dimensional space.

4.4 Comparisons

To evaluate how well TAWE assigns titles to topics, we need a gold standard that allows us to understand if our method is performing better or worse than other existing methods.

Inspired by the approach of Lau et al. (2011), we chose to use the Amazon Turk platform to recruit human evaluators who expressed their preference for some proposed titles assigned to topics. A topic for each newspaper was randomly selected. Therefore, users expressed their opinion on a selection of six topics, and the experiment was set up in such a way that, for each topic, the list of the first $M = 100$ words was provided together with three potential titles: (1) TAWE's title: the title attributed by TAWE; (2) HiProb's title: the title attributed based on the word most associated with the topic in question; (3) Chat GPT's title: the title attributed by Chat GPT providing the list of $M = 100$ words and with the following instruction: "Suggest a title (a single word) for each one of these lists". Each title was evaluated by 50 raters with a score ranging from 1 (minimum) to 5 (maximum). The options were presented randomly to avoid order bias. In Table 8, we report the average rating given by Amazon Turk users.

Table 5 Top ten words for topics identified in The New York Times tweets

Topic	Top ten words
Topic 1	study (0.017), water (0.014), found (0.011), levels (0.011), catastrophic (0.009) glaciers (0.009), home (0.009), worse (0.009), coal (0.009), experts (0.008)
Topic 2	president (0.024), americans (0.019), summit (0.018), leaders (0.017), flooding (0.015) biden (0.013), natural (0.013), scientists (0.012), year (0.011), action (0.010)
Topic 3	trump (0.029), administration (0.019), president (0.013), life (0.010), political (0.009) part (0.009), report (0.009), white (0.009), fires (0.009), back (0.008)
Topic 4	weather (0.019), temperatures (0.015), future (0.015), extreme (0.011), storms (0.010) rising (0.010), thursday (0.009), areas (0.009), local (0.009), vulnerable (0.009)
Topic 5	people (0.013), action (0.010), government (0.010), effects (0.010), species (0.009) fighting (0.009), report (0.009), long (0.009), policy (0.009), power (0.008)
Topic 6	gas (0.016), fossil (0.016), fuel (0.016), emissions (0.014), companies (0.011) west (0.010), record (0.010), biggest (0.010), fire (0.009), exxon (0.008)
Topic 7	biden (0.048), president (0.032), countries (0.016), plan (0.011), news (0.011) joe (0.011), major (0.010), breaking (0.010), people (0.009), democrats (0.009)
Topic 8	president (0.020), energy (0.014), trump (0.013), bill (0.011), city (0.011) time (0.010), house (0.008), fight (0.008), renewable (0.008), bidens (0.007)
Topic 9	china (0.017), jobs (0.014), economic (0.011), work (0.011), residents (0.010) issues (0.010), forced (0.009), huge (0.008), long (0.008), told (0.008)
Topic 10	year (0.014), effects (0.012), federal (0.011), opinion (0.010), times (0.010) york (0.009), make (0.008), day (0.007), scientists (0.007), gas (0.007)

The title provided by TAWE received the best score in four cases out of six, and second-best after Chat GPT in the remaining two cases. The title assigned by Chat GPT obtained the best score in two cases out of six and the second-best score in the remaining four cases. Based on the obtained performance of Chat GPT titles according to the evaluation of Amazon Turk raters, we decided to use Chat GPT as the gold standard in the comparison between TAWE

Table 6 Top ten words for topics identified in The Washington Post tweets

Topic	Top ten words
Topic 1	biden (0.037), analysis (0.022), trump (0.022), administration (0.019), energy (0.018) president (0.015), fight (0.012), action (0.011), finds (0.011), study (0.010)
Topic 2	trump (0.015), summit (0.013), leaders (0.013), people (0.012), perspective (0.012) opinion (0.012), activists (0.011), post (0.010), scientists (0.009), hurricanes (0.009)
Topic 3	house (0.027), white (0.023), epa (0.011), blame (0.010), people (0.009) greenhouse (0.008), report (0.008), young (0.008), isn (0.008), agency (0.008)
Topic 4	post (0.016), nations (0.014), discuss (0.014), united (0.012), time (0.011) analysis (0.011), summer (0.011), scientists (0.010), slow (0.009), stories (0.008)
Topic 5	carbon (0.020), heat (0.016), opinion (0.016), fight (0.015), weather (0.015) make (0.012), human (0.012), cut (0.011), scientists (0.011), analysis (0.011)
Topic 6	opinion (0.015), oil (0.010), analysis (0.009), companies (0.009), francis (0.009) reports (0.009), growing (0.009), energy (0.009), pope (0.009), threat (0.008)

Table 7 Top ten words for topics identified in The Wall Street Journal tweets

Topic	Top ten words
Topic 1	president (0.013), trump (0.011), energy (0.011), biden (0.010), writes (0.007)
	presidential (0.006), renewable (0.006), make (0.006), joe (0.006), impact (0.005)
Topic 2	energy (0.014), renewable (0.009), financial (0.009), companies (0.008), president (0.007)
	york (0.007), writes (0.006), deal (0.005), cost (0.005), electric (0.005)
Topic 3	country (0.011), health (0.011), beijing (0.009), coffee (0.009), energy (0.009)
	chinese (0.008), renewable (0.007), businesses (0.007), fight (0.006), cost (0.006)
Topic 4	biden (0.010), china (0.009), president (0.009), address (0.008), companies (0.006)
	impact (0.006), calling (0.006), joe (0.006), opinion (0.005), congress (0.005)
Topic 5	energy (0.016), power (0.013), renewable (0.012), gas (0.011), oil (0.010)
	government (0.008), obama (0.007), coal (0.007), top (0.006), fight (0.006)
Topic 6	emissions (0.011), biden (0.010), china (0.010), energy (0.007), gas (0.007)
	trump (0.007), president (0.007), companies (0.006), investors (0.006), epa (0.006)

and the HiProb's title (the title assigned based on the probabilities obtained when creating topics with LDA, or with other topic modeling techniques, and selecting only the first word) assigned in the totality of topics of this application as well as in the second application (described in the Supplementary material). To this aim, the cosine similarity between these two titles and the title provided by Chat GPT was computed through the cosine function of the lsa R package (Wild, 2007). A higher value of cosine similarity, which ranges from 0 (minimum) to 1 (maximum), was considered to indicate the best title. For each topic, besides the Chat GPT title assigned as previously described, we also generated a second Chat GPT

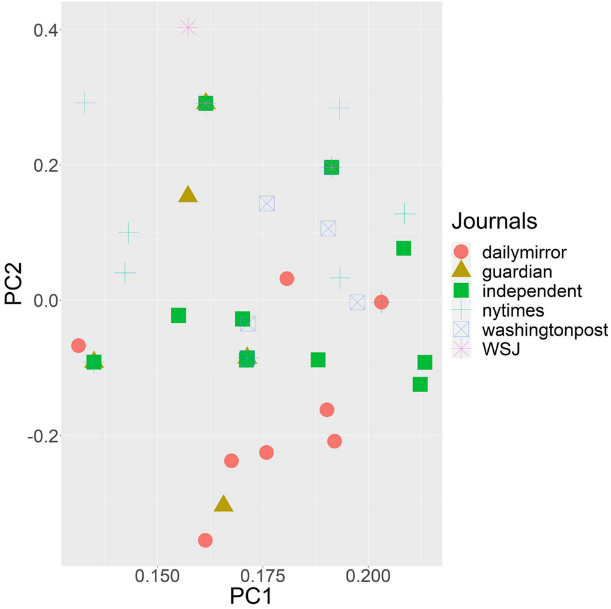

Fig. 6 Representations with principal components of all topics labels for all newspapers

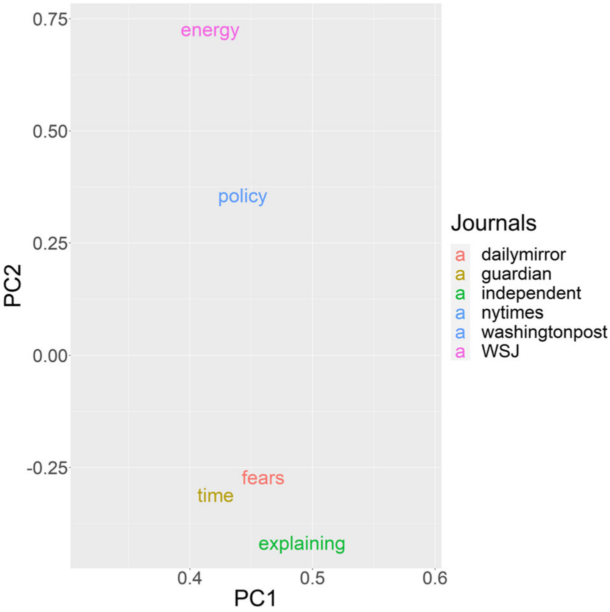

Fig. 7 Representations with principal components of all centroid labels for all newspapers

title for which the query entered into the system ("Choose the most representative word for each one of these lists") required the choice to be limited to the 100 provided words (restricted version).

In Table 9, we report the comparisons between the titles assigned by the TAWE and the HiProb methods with respect to the title provided by Chat GPT. As the table shows, in 40 out of 50 cases (for the first set of comparisons) and in 26 out of 50 cases (for the second set of comparisons), TAWE has a value greater than or equal to that of HiProb, and this means that the title assigned by TAWE is more often closer to that assigned by Chat GPT than the title assigned by HiProb. In a second application (reported in full in the Supplementary material) based on the same data set but where topics were obtained using a different technique (STM) the results were very similar to this first application, where TAWE performed better than the HiProb method. We presented this second application with the aim of confirming that TAWE

Table 8 Rating of TAWE, HiProb, and Chat GPT using Amazon Turk's crowdsourcing service

Topic	Newspaper	TAWE Title	Rating	HiProb Title	Rating	Chat GPT Title	Rating
1	GUA	time	2.18	view	2.56	coverage	**3.60**
9	DM	threatening	**3.76**	facing	2.32	solutions	2.44
18	IND	dangers	**3.58**	scientists	2.96	public	2.60
38	NYT	concerns	3.84	year	2.00	crisis	**4.10**
39	WP	policy	**3.60**	biden	2.56	analysis	2.98
47	WSJ	energy	**3.12**	country	3.10	china	2.84

Abbreviations: *GUA*, The Guardian; *DM*, Daily Mirror; *IND*, The Independent; *NYT*, The New York Times; *WP*, The Washington Post; *WSJ*, The Wall Street Journal

The best results are reported in bold

Table 9 Comparisons of TAWE and HiPRob versus Chat GPT using cosine similarity

TAWE	HiProb	ChatGPT	TAWE vs ChatGPT	HiProb vs ChatGPT	ChatGPT (R)	TAWE vs ChatGPT (R)	HiProb vs ChatGPT (R)
The Guardian							
Time	View	Coverage	**0.49**	0.49	Crisis	**0.40**	0.35
Trump	Trump	Implications	**0.23**	0.23	Court	**0.43**	0.43
Energy	Energy	Challenges	**0.41**	0.41	Renewable	**0.81**	0.81
Study	Study	Findings	**0.67**	0.67	Study	**1.00**	1.00
Fight	Fight	Consequences	**0.43**	0.43	Leaders	**0.41**	0.41
The Daily Mirror							
Threat	Experts	Urgency	**0.57**	0.39	Experts	0.42	1.00
Promises	Johnson	Impact	**0.38**	0.31	Plastic	0.25	0.30
Find	Find	Awareness	**0.32**	0.32	Impact	**0.34**	0.34
Threatening	Facing	Solutions	**0.35**	0.28	Energy	**0.31**	0.22
Furious	Forever	Threats	**0.42**	0.31	Risk	**0.28**	0.27
Threat	Fight	Adaptation	**0.40**	0.36	Sea	**0.38**	0.27
Time	Day	Warnings	0.26	0.36	Day	0.66	1.00
Fears	Donald	Trump	0.32	0.60	Donald	0.31	1.00
Wild	Humans	Environment	0.28	0.51	Humans	0.47	1.00
Due	Act	Climate	**0.35**	0.20	Act	0.34	1.00
Fears	Prince	Leadership	**0.43**	0.32	Prince	0.29	1.00
Frightening	Brits	Effects	**0.51**	0.28	Brits	0.30	1.00
The Independent							
Dangers	Scientists	Public	**0.32**	0.24	Scientists	0.53	1.00
Protest	Report	Reports	0.41	0.83	Report	0.48	1.00
Policy	Children	Policies	**0.88**	0.32	Children	0.30	1.00
Opinion	Ppinion	Debates	**0.67**	0.67	Opinion	**1.00**	1.00
Time	Fight	Denial	0.35	0.44	Energy	**0.41**	0.21

Table 9 continued

TAWE	HiProb	ChatGPT	TAWE vs ChatGPT	HiProb vs ChatGPT	ChatGPT (R)	TAWE vs ChatGPT (R)	HiProb vs ChatGPT (R)
Energy	Energy	Energy	**1.00**	1.00	Watch	**0.32**	0.32
Shows	Watch	Data	**0.39**	0.34	House	**0.36**	0.36
Fears	Attenborough	Advocacy	**0.35**	0.24	Attenborough	0.31	1.00
Bringing	Plastic	Accord	**0.39**	0.13	Leaders	**0.45**	0.05
Trump	Trump	Denier	**0.34**	0.34	Denier	**0.34**	0.34
Explaining	Study	Evidence	**0.55**	0.54	Study	0.53	1.00
The New York Times							
Impacts	Study	Catastrophic	**0.66**	0.32	Levels	**0.55**	0.46
Threat	President	Action	**0.48**	0.23	Summit	0.36	0.38
Trump	Trump	Administration	**0.43**	0.43	Administration	**0.43**	0.43
Weather	Weather	Disasters	**0.47**	0.47	Temperatures	**0.55**	0.55
Policy	People	Effects	**0.41**	0.30	Effects	**0.41**	0.30
Gas	Gas	Emissions	**0.67**	0.67	Emissions	**0.67**	0.67
Biden	Biden	Trillion	**0.35**	0.35	President	**0.58**	0.58
Time	President	Climate	**0.33**	0.19	President	0.34	1.00
Efforts	China	Existential	**0.29**	0.19	Fossil	0.21	0.35
Concerns	Year	Crisis	**0.57**	0.31	Year	0.27	1.00
The Washington Post							
Policy	Biden	Analysis	**0.55**	0.29	Analysis	**0.55**	0.29
Leaders	Trump	Summit	**0.47**	0.42	Trump	0.36	1.00
Experts	House	Policies	**0.54**	0.26	House	0.28	1.00
Time	Post	Perspectives	0.30	0.34	Weather	**0.42**	0.31
Focus	Carbon	Impact	**0.65**	0.49	Carbon	0.36	1.00
Future	Opinion	Opinion	0.39	1.00	Opinion	0.39	1.00

Table 9 continued

TAWE	HiProb	ChatGPT	TAWE vs ChatGPT	HiProb vs ChatGPT	ChatGPT (R)	TAWE vs ChatGPT (R)	HiProb vs ChatGPT (R)
The Wall Street Journal							
Policy	President	President	0.46	1.00	President	0.46	1.00
Energy	Energy	Energy	**1.00**	1.00	Energy	**1.00**	1.00
Energy	Country	China	0.26	0.40	Country	0.27	1.00
Policy	Biden	Biden	0.46	1.00	Biden	0.46	1.00
Energy	Energy	Power	**0.65**	0.65	Energy	**1.00**	1.00
Emissions	Emissions	Emissions	**1.00**	1.00	Emissions	**1.00**	1.00

The best results are reported in bold

can be applied to topics created with any technique and that it can provide good titles in an automated fashion.

5 Conclusions

We introduced a new method called TAWE to automatically assign titles to topics. Starting from a topic modeling method (e.g., LDA), we extracted the most important words for each topic and their weights. Next, we used word embedding to transform such weighted words into vectors and to find the centroid for all vectors. Using cosine similarity, we looked for the word closest to the centroid. This word is selected as the most representative word of the set. TAWE allows to avoid the subjectivity of the manual attribution of titles to topics, which is a job often carried out by researchers.

An application of the TAWE method to a data set of tweets about climate change, and in which topics were obtained with LDA, was described in Sect. 4. However, the described TAWE method can also be applied when the topics are obtained with a different topic modeling method, as shown in a second application described in the Supplementary material. In both applications the TAWE method achieved a good performance in the automatic assignment of titles. In addition, through a bidimensional representation of topics, we showed how TAWE can allow to assess similarities and differences between sources (e.g., newspapers) as regards to topics covered in climate change tweets.

There are few limitations to this procedure. Firstly, the number of words fed to the TAWE algorithm is a parameter that has to be decided by the researcher. We showed in Sect. 4.3 that the number of words used to assign titles affected the results since the more words we use, the more noise we introduce into the procedure. Secondly, n-grams cannot be used in the procedure without considering an ad hoc created word embedding that relays in a loss of information in the definition of titles. Future research will be aimed at implementing the ability to deal with n-grams.

Appendix

Table 10 Perplexity and coherence scores for each newspaper for 2 to 20 topics

	GUA		DM		IND		NYT		WP		WSJ	
#	P	C	P	C	P	C	P	C	P	C	P	C
2	−13.33	0.54	−7.44	0.64	−14.01	0.52	−11.35	0.37	−11.43	0.43	−8.30	0.43
3	−13.28	0.56	−7.31	0.58	−13.96	0.50	−11.32	0.33	−11.41	0.46	−8.29	0.44
4	−13.27	0.56	−7.27	0.54	−13.91	0.54	−11.31	0.37	−11.40	0.43	−8.28	0.44
5	−13.27	0.60	−7.21	0.55	−13.88	0.56	−11.29	0.37	−11.39	0.49	−8.28	0.45
6	−13.27	0.58	−7.20	0.53	−13.86	0.55	−11.29	0.39	−11.38	0.54	−8.29	0.48
7	−13.25	0.60	−7.17	0.48	−13.85	0.56	−11.27	0.41	−11.40	0.51	−8.26	0.46
8	−13.24	0.59	−7.15	0.47	−13.82	0.57	−11.27	0.43	−11.41	0.53	−8.30	0.48
9	−13.26	0.61	−7.13	0.44	−13.83	0.58	−11.29	0.41	−11.40	0.53	−8.29	0.46
10	−13.26	0.62	−7.14	0.43	−13.80	0.57	−11.29	0.45	−11.42	0.55	−8.28	0.45
11	−13.26	0.63	−7.11	0.38	−13.80	0.60	−11.29	0.46	−11.43	0.54	−8.29	0.44

Table 10 continued

	GUA		DM		IND		NYT		WP		WSJ	
#	P	C	P	C	P	C	P	C	P	C	P	C
12	−13.27	0.64	−7.11	0.41	−13.81	0.60	−11.29	0.47	−11.42	0.54	−8.31	0.45
13	−13.28	0.62	−7.11	0.41	−13.81	0.61	−11.29	0.48	−11.46	0.57	−8.34	0.46
14	−13.28	0.61	−7.10	0.41	−13.80	0.61	−11.30	0.48	−11.45	0.56	−8.32	0.47
15	−13.29	0.65	−7.11	0.37	−13.81	0.61	−11.32	0.49	−11.47	0.55	−8.34	0.46
16	−13.30	0.63	−7.11	0.38	−13.81	0.61	−11.31	0.48	−11.46	0.58	−8.36	0.49
17	−13.31	0.64	−7.12	0.39	−13.85	0.64	−11.32	0.48	−11.48	0.57	−8.36	0.46
18	−13.32	0.65	−7.14	0.39	−13.83	0.62	−11.33	0.49	−11.50	0.59	−8.37	0.45
19	−13.35	0.66	−7.14	0.35	−13.85	0.63	−11.33	0.48	−11.50	0.56	−8.36	0.45
20	−13.34	0.66	−7.16	0.40	−13.84	0.63	−11.34	0.51	−11.52	0.59	−8.36	0.46

Abbreviations: *Gua*, The Guardian; *DM*, Daily Mirror; *Ind*, The Independent; *NYT*, The New York Times; *WP*, The Washington Post; *WSJ*, The Wall Street Journal; #, number of topics; *P*, perplexity; *C*, coherence score

Table 11 Labels for all 10 topics identified in The Guardian while varying $m = 10, \ldots, 100$

m	var	median	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
10	0.0107	0.3147	View	Trump	Renewable	Study	Fight
20	0.0145	0.3045	Policy	Trump	Energy	Study	Car
30	0.0120	0.3126	Policy	Trump	Energy	Study	Black
40	0.0180	0.3212	Time	Trump	Energy	Study	Emissions
50	0.0066	0.3229	Time	Trump	Energy	Study	Make
60	0.0092	0.3666	Time	Planet	Energy	Study	Scientists
70	0.0033	0.3796	Time	Ignore	Energy	Study	Weather
80	0.0226	0.2937	Time	Gas	Energy	Study	Weve
90	0.0047	0.4008	Time	Scientist	Energy	Study	Project
100	0.0057	0.3826	Time	Extinction	Energy	Study	Coral

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement. This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP F53C22000760007, GZ and CC) and the e.INS – Ecosystem of Innovation for Next Generation Sardinia project (e.INS ECS00000038 - CUP F53C22000430001, MR and CC). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Data Availability The data and the R code that support the findings of this study are available on GitHub repository: https://github.com/z-gp/tawe

Declarations

Ethics Approval This research does not contain any studies with human participants or animals.

Conflict of Interest The authors declare no competing interests.

References

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2000). A neural probabilistic language model. *Advances in neural information processing systems, 13*.

Bhatia, S., Lau, J. H., & Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers* (pp. 953–963). Osaka, Japan. The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1091

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. L. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Calleo, Y., & Pilla, F. (2022). Using geo-spatial topic modelling to understand the public view of Italian Twitter users: a climate change application. *SIS 2022 Proceedings*.

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. H. (2010). Short and tweet: Experiments on recommending content frominformation streams. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1185–1194).

Dahal, B., Kumar, S. A. P., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining, 9*(1), 1–20.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

DePaula, N. (2020). Climate science communication on twitter: A topic modeling analysis of US federal government agencies. *iConference 2020 Proceedings*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805

Dumanovsky, T., Huang, C. Y., Bassett, M. T., & Silver, L. D. (2010). Consumer awareness of fast-food calorie information in New York City after implementation of a menu labeling regulation. *American Journal of Public Health, 12*, 2520–2525.

Effrosynidis, D., Karasakalidis, A. I., Sylaios, G., & Arampatzis, A. (2022). Controversy around climate change reports: A case study of Twitter responses to the 2019 IPCC report on land. *Expert Systems with Applications, 204*, 117541.

Felaco, C., Mazza, R., & Parola, A. (2020). A mixture of topic modeling and network analysis. the case-study of climate change on Twitter. *CCSD (Vorsitz), 15es Journées internationales d'Analyse statistique des Donnés Textuelles, Toulouse*.

Gallagher, R. J., Reing, K., Kale, D., & Ver Steed, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. In *2017 Transactions of the Association for Computational Linguistics (TACL)* (pp. 529–538).

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science advances, 4*(7), eaaq1360.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. *Proceedings of the 24th annual conference of the cognitive science society* (pp. 381–386).

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International SIGIR conference on research and development in information retrieval* (pp. 50–57).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning,* pp. 177–196.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks, 5*(2), 109–137.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89–106.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169–15211.

Jeong, B., Yoon, J., & Leeb, J.-M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management, 48*, 280–290.

Jin, X., & Han, J. (2010). *K-Medoids Clustering* (pp. 564–565). US, Boston, MA: Springer. ISBN 978-0-387-30164-8. https://doi.org/10.1007/978-0-387-30164-8_426

Killick, R., & Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software, 58*, 1–19.

Kozono, R., & Saga, R. (2020). Automatic labeling for hierarchical topics with NETL. In: *2020 IEEE International conference on Systems, Man, and Cybernetics (SMC)* (pp. 3740–3745). https://doi.org/10.1109/SMC42975.2020.9282874

Landauer, T. K., & Dumais, S. T. (1997). Asolution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line trend analysis with topic models: Twitter trends detection topic model online. *Proceedings of COLING 2012* (pp. 1519–1534).

Lau, J. H., Grieser, K. , Newman, D. & Baldwin, T. (2011). Automatic labelling of topic models. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 1536–1545). Portland, Oregon, USA. Association for Computational Linguistics. https://aclanthology.org/P11-1154

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*.

Lee, C., & Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science, 4*(1), 1–50.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning* (pp. 577–584).

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *Springer Plus, 5*(1), 1–22.

Mabey, B. (2021). pyLDAvis 2.1.2 documentation. https://pyldavis.readthedocs.io/en/latest/. Accessed 30 Oct 2022.

Marlow, T., Miller, S., & Roberts, T. (2021). Bots and online climate discourses: Twitter discourse on president Trump's announcement of US withdrawal from the paris agreement. *Climate Policy, 21*(6), 765–777.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111–3119.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 551–554.

Ohtani, S. (2022). How is people's awareness of biodiversity measured? using sentiment analysis and lda topic modeling in the twitter discourse space from 2010 to 2020. *SN Computer Sciencey, 3*(371).

R Core Team (2020). R: A language and environment for statistical computing. r foundation for statistical computing, Vienna, Austria v. 4.1.2. http://www.R-project.org.

Řehůřek, R., & Sojka, P. (2011). Gensim, statistical semantics in python.

Sanford, M., Painter, J., Yasseri, T., & Lorimer, J. (2021). Controversy around climate change reports: a case study of twitter responses to the 2019 IPCC report on land. *Climatic Change, 167*(3), 1–25.

Scott, A. J., & Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics, 30*, 507–512.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Sindhu, K., & Seshadri, K. (2021). Dimensionality prediction for word embed dings. In: X.-Z. Gao, R. Kumar, S. Srivastava, & B. P. Soni (Eds.), *Applications of artificial intelligence in engineering* (pp. 301–317). Singapore, Springer Singapore. ISBN 978-981-33-4604-8. https://doi.org/10.1007/978-981-33-4604-8_24.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267–276.

Truică, C.-O., & Apostol, E.-S. (2021). TLATR: Automatic topic labeling using automatic (domain-specific) term recognition. *IEEE Access, 9*, 76624–76641. https://doi.org/10.1109/ACCESS.2021.3083000

Van Lange, P. A. M., & Huckelba, A. L. (2021). Psychological distance: How to make climate change less abstract and closer to the self. *Current Opinion in Psychology, 42*, 49–53.

Vavliakis, K. N., Symeonidis, A. L., & Mitkas, P. A. (2013). Consumer awareness of fast-food calorie information in New York City after implementation of a menu labeling regulation. *Data & Knowledge Engineering, 88*, 1–24.

Wakefield, M., Flay, B., Nichter, M., & Giovino, G. (2003). Role of the media in influencing trajectories of youth smoking. *Addiction, 98*, 79–103.

Wild, F. (2007). An LSA package for R. In: *Proceedings of the 1st international conference on Latent Semantic Analysis in Technology Enhanced Learning (LSATEL'07)* (pp. 11–12).

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science,2*(2), 165–193. ISSN 2198-5812. https://doi.org/10.1007/s40745-015-0040-1

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 23–30). Association for Computational Linguistics.

Zammarchi, G., Romano, M., & Conversano, C. (2022). Evolution of media coverage on climate change and environmental awareness: An analysis of tweets from UK and US newspapers. *Classification and Data Science in the Digital Age - Book of Abstracts IFCS 2022* (p. 122).