Towards a Transparent and an Environmental-Friendly Approach for Short Text Topic Detection: A Comparison of Methods for Performance, Transparency, and Carbon Footprint

Sami Al Sulaimani* and Andrew Starkey
School of Engineering, University of Aberdeen, Aberdeen AB24 3UE, UK
*Correspondence: s.alsulaimani1.19@abdn.ac.uk (S.A.S.)

***Abstract*—Online social media platforms have contributed significantly to the dissemination of user-generated information. Many studies have proposed various techniques to analyze publicly available short texts to automatically extract topics. The majority of these works have mainly focused on the competitive performance of the proposed approaches. In this paper, our main focus is on how to tackle this problem by incorporating two other important qualities: Transparency and Carbon Footprint. These two pillars are cornerstones to fulfill the emerging international demands and to adhere to the new regulations, such as "Right to Explanation" and "Green AI". Based on these three qualities, this paper compares the most prominent algorithms in this field (specifically within the category of unsupervised-retrospective learning), such as: Latent Dirichlet Allocation, Non-Negative Matrix Factorization, and K-Means, as well as two most recent approaches, such as: BERTopic and Contextual Analysis. By using two different datasets, the methods were evaluated for Performance. On average, the results show that BERTopic is the best-performing approach overall in terms of Performance. However, Contextual Analysis achieves the best Performance in one of the two datasets used. When considering the three qualities together, the results demonstrate the effectiveness and the benefits of the Contextual Analysis method towards a more transparent and greener approach for the topic detection task.**

***Keywords*—text analysis, topic detection, contextual analysis, unsupervised machine learning, carbon footprint, transparency, explainability**

I. Introduction

The increase of user-generated information has attracted the attention of many researchers and practitioners all over the world. There have been various exciting studies in which many tools and solutions have been proposed in order to leverage the availability of the disseminated contents (i.e., texts, images, videos) to extract knowledge for various useful applications. One interesting research direction focuses on automatically discovering topics from text contents [1].

The introduction of online social media platforms, such as Twitter, YouTube, and Facebook, has contributed significantly to this growth. People use these very accessible and easy-to-use facilities to share a wide range of their daily life topics. According to [2], the number of reported worldwide social media users in January 2022 is around 4.6 billion, a growth of 10% over 12 months. This has led to a growing opportunity for the emergence of new applications that can harness these contents, most importantly when there are no alternative sources for this information [3].

This evolution has brought new challenges to the traditional topic detection task [4], where the goal was to identify topics from long and relatively well-written texts (such as academic papers and news web pages). The text on prominent social media platforms is short. For example, Twitter allows a maximum of 280 characters in the user posts [5]. In these channels, users tend to extensively use slang words, uncommon abbreviations, emoticons, and misspelled words. in their texts. These challenges have negatively impacted the performance of the classical approaches, which in turn have motivated the community to develop new techniques and methods in order to overcome them.

Lately, researchers have proposed many useful topic detection frameworks from social media for various applications, such as disease outbreak detection [6], disaster management [7], riot control [8] and crime and terrorism prevention [9]. For example, researchers in [10] analyzed five different machine learning models for cyberbullying detection using a Twitter dataset, focusing on text features. Furthermore, the authors in [11] employed both supervised and unsupervised machine learning techniques to detect topics from citizen complaints related to government services, such as floods and damaged roads.

These approaches can generally be categorized into various groups, such as online or retrospective, supervised or unsupervised, and with or without neural word embedding[1] (such as word2vec and BERT) [1], [13], [4], [14]. While the online approaches ([15], [16], [17]) focus on extracting topics from real-time posts (i.e., as soon as posts arrive), the retrospective solutions ([18], [19]) receive a whole corpus (i.e., collections of recorded posts in the past) as an input for analysis in an offline manner. In the supervised approaches ([20], [21]), a labeled dataset is used to train a classifier which will be used for the detection purpose. On

[1]Neural word embedding is one type of word representation that has been widely employed in solving various text mining problems. It is an effort that aims to represent words as low-dimensional vectors to store their contextual information, in which similar words (i.e., used in a similar way) have similar representations [12].

Manuscript received June 7, 2023; revised July 7, 2023; accepted July 14, 2023; published November 22, 2023.

the other hand, the unsupervised solutions ([22], [23]) do not require the labeled dataset to achieve the topic detection task. Other hybrid approaches, like semi-supervised or online-offline, have been proposed in the literature.

Whilst the topic detection problem has been extensively studied in previous works, less attention has been devoted to the transparency and the environmental-friendly aspects of the task. Approaches based on deep learning methods, for instance, can use significant computational resources and are based on complex internal structures (e.g., training a BERT model in a graphics processing unit (GPU) consumes about 1,507 kWh [24]). These qualities are essential to answer the emerging international demands and regulations [25], [26], [27], such as "Right to Explanation" and "Green AI", which are fully discussed in subsequent sections. To the best of our knowledge, no study on the topic detection approaches in terms of the three qualities (see Fig. 1), i.e., Performance, Transparency, and Carbon Footprint has been conducted. To this extent, and complementing the existing works, the primary objective of this paper is to justify the need for such methods to solve this task. Then, to fill the gap, it compares key algorithms in this field, such as: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and K-Means, as well as two more recent approaches, namely BERTopic and Contextual Analysis (CA) and compares them in terms of their performance against these three qualities. Also, this is the first study that examines the capabilities of Contextual Analysis in an unsupervised-retrospective topic detection task. Note: readers interested in exploring a comparative analysis of various key supervised learning methods, in terms of Performance and Explainability, are referred to our recent research, which can be found in [28].

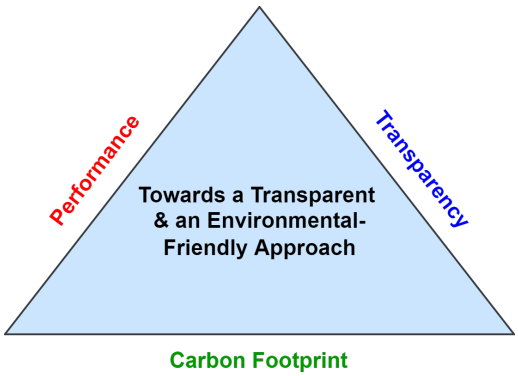

Figure 1. A triangle representing three qualities that highlight the scope of the review: Performance, Carbon Footprint, and Transparency.

The rest of this paper is organized as follows. Section II briefly reviews the related works. Section III justifies the need for environmental-friendly approaches. Section IV highlights the importance of transparent algorithms. Section V describes our survey to explore the existence of Transparency or Carbon Footprint assessments in the published topic detection approaches. Section VI discusses experiments and results obtained by comparing five approaches based on three qualities: Performance, Carbon Footprint, and Transparency. Finally, Section VII gives discussion and concluding remarks as well as our recommendations for future work.

II. Related Work

In the last years, several surveys in the field of topic detection have been presented in the literature. While some efforts explored the various proposed methods, others focused on other related problems, like the evaluation techniques (i.e., how to compare methods against each other?).

In [29], the authors reviewed various evaluation techniques for the detection task for Twitter, in which run-time and task-based measures were proposed. The first focuses on assessing the methods based on the number of processed tweets per second and the amount of utilized memory. In the second, three metrics were presented, including: duplicate detection rate, precision, and recall, in which other external resources (such as Google Search, New York Times archive, and Reuters website) were utilized. Then, based on these metrics, nine of the detection approaches were assessed. However, the dependency on external sources is the major disadvantage of the proposed measures: (1) issues such as changes in the sources' internal policies and restrictions may significantly affect the evaluation, (2) local or sub-topics may not be available in these services.

The efforts in [1] surveyed the event detection based on Twitter from different perspectives, such as: detection approaches, data collection, evaluation strategies, limitations, and research trends. Also, a taxonomy based on "event type", "detection approach", "orientation", and "application domains" for the detection approaches was presented. The authors found that precision, recall, f1, accuracy, and error rate are the widely used evaluation metrics. They argued that the selection of any measure and the distribution of the classes should be well considered. Although the work highlighted some important challenges and provided interesting recommendations, it did not present any experimental analysis of the detection techniques.

The work presented by Nugroho *et al.* [30] reviewed the detection approaches on social media (mainly Twitter), with more focus on the features they used, including: content (e.g., tweet texts), social interactions (e.g., mentions, retweets, hashtags) and temporal (e.g., tweet's arrival time). Also, an experimental study was conducted to compare the performance of the most prominent algorithms (and their variants) in the field, for instance, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorisation (NMF) and their extensions. They found that adding more features, such as social interactions and temporal, improved the performance. However, one should note that the high dependency of methods on specific attributes (that are provided by certain platforms, like Twitter) may limit their usage.

Mottaghinia *et al.* [13] categorized the reviewed topic detection approaches in Twitter into four categories: with or without word embedding (i.e., whether it employs pre-trained models, like word2vec, BERT, or not), specified or unspecified (i.e., whether it depends on prior information, like the place, time, or not), offline or online (i.e., whether it receives inputs as a whole corpus or as a real-time data stream) and supervised or unsupervised (i.e., whether it

requires labeled training data or not). The authors found that the majority of detection approaches are clustering-based techniques.

More recently, Tijare and Rani Prathuri [14] placed particular emphasis on offline and online event detection approaches in social media. They found that most of the reviewed studies fall under the offline category. Moreover, a list of datasets that were used in various studies in the field is presented. For instance, EVENT2012 is one of the corpuses that were used by some of the reviewed studies.

To the best of our knowledge, none of the existing works have discussed the Transparency and the Carbon Footprint aspects of the topic detection task. Therefore, differently from the existing reviews; however complementing them, this paper reviews the topic detection approaches based on three qualities: Performance, Carbon Footprint, and Transparency.

III. The Need for Environmental-Friendly Approaches

With the advanced capabilities of computational resources, researchers have managed to develop complex algorithms that can produce unprecedented results across various tasks. Depending on the problem at hand, some algorithms utilize enormous resources for days or months to complete their task or to achieve competitive performance. For example, training the NAS model on 8 NVIDIA P100 GPUs requires 274,120 compute hours and consumes about 656,347 kWh of energy [24]. This is equivalent to the consumption of 226 domestic electricity meters in the UK per year (Note: in the UK, the median domestic electricity consumption is 2,902 kWh per year [31]). This has brought attention to the tremendous energy that is required to run these approaches and to the study of their negative impact on the environment due to the produced greenhouse gas emissions.

The impact of greenhouse gas emissions and the demand for urgent actions to counter their accelerated threats on the earth (such as: human health, poverty, rise in sea level, drought, and species loss) have motivated various international events and assemblies (more recently the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow). According to The Intergovernmental Panel on Climate Change (IPCC), it is projected that 1.5 degrees Celsius of global warming can be reached between 2030 and 2052 [32]. To limit this accelerated change, governments have entered into international agreements, like the Paris Agreement [27], and have committed to international roadmaps, such as to achieve "energy-related carbon dioxide emissions to net zero by 2050" [33]. This commitment requires countries to report their Nationally Determined Contributions (NDCs), which is "a climate action plan to cut emissions and adapt to climate impacts" [34]. Mitigating the serious climate change-related risks requires ever-increasing international actions and multidisciplinary participation by academics and practitioners from various fields.

In light of this, the study of the environmental cost of Artificial Intelligence approaches, more specifically in machine learning algorithms, has been a trending research line and has brought the attention of many researchers and practitioners in the community [24], [35], [36], [37] [38], [39]. In [24], the authors proposed a method to quantify the environmental costs of training neural network models for NLP. By measuring the power that is consumed by CPU, DRAM, GPU, and Power Usage Effectiveness (PUE) during the training phase, they compute the Carbon Footprint (i.e., an approximation of the carbon dioxide equivalent (CO_2eq) emissions which is "a measure of how much a gas contributes to global warming, relative to carbon dioxide" [40]). The authors reported that training the NAS model (using the hardware mentioned above) is estimated to contribute 626k pounds of CO_2eq emissions into the environment, which is roughly equivalent to the carbon sequestered by 336 acres of U.S. forests in one year (using Greenhouse Gas Equivalencies Calculator [41]).

It is worth noting that several efforts have been made to outline strategies in order to mitigate the impacts of the algorithms on the environment [24], [37], [36], [42]. Computing and declaring the Carbon Footprint related metric is the common recommendation among them (which was not recorded in the published papers according to the study conducted by Henderson *et al.* [36] on a sample of 100 NeurIPS papers from the 2019 proceedings). Bommasani *et al.* [42] suggested some practices that can be useful in this regard, such as: choosing low-carbon intensity regains to train models, selecting a model (small vs. large) should carefully consider its costs and benefits to society, and reporting the energy, computational, and carbon costs of the used model. Also, in [24], the authors advocate the encouragement of computationally efficient algorithms and hardware, as well as the importance of reporting a model's training time, computational resources, and sensitivity to hyperparameters.

To facilitate the quantification of the Carbon Footprint, researchers have offered some easy-to-use techniques and tools that can be employed during the computational task. Lacoste *et al.* [35] proposed an approach called the Machine Learning Emissions Calculator. By receiving four inputs, including: hardware type, training duration, cloud service provider, and training region, it estimates the carbon emissions. More recently, other tools like the ones in [36] (called Experiment Impact Tracker), in [38] (called CarbonTracker), in [43] (called CodeCarbon) and in [44] (called Green Algorithms) have been proposed.

IV. The Need for Transparent Approaches

The growing complexity of machine learning algorithms raises other concerns that are related to their opaque nature and the lack of transparency. Many AI approaches, for instance, the ones that fall under the deep learning umbrella, can achieve impressive results in terms of their Performance (i.e., classification accuracy); however, this comes at the cost of complex internal interactions that cannot be directly understood [45], [46]. This opacity has become a growing concern and has triggered the need to clearly understand the automated decisions that are made by these methods, more importantly, if they are intended for use in some critical or highly sensitive domains, such as: health (emergency triage), transportation (automated vehicle), finance (credit scoring), human resource (hiring), Justice (criminal justice), public safety (terrorist detection), and so on. The question

that arises here is to which degree can we trust (or fully rely on) the algorithmic decisions that can be biased or erroneous and complex or opaque (difficult to comprehend)?

Justifying algorithmic outputs, more importantly, when something goes wrong, is one of many reasons why we need to open the black-box. Having the ability to clearly understand the outputs and decision-making process made by the AI can help to improve the methods towards better outcomes [45]. Also, with more information, rather than just one output, this can offer other ways to explore data and reveal unknown knowledge and insights. However, one should also note that providing too many details can negatively impact the understandability of the method.

Toward addressing the raised issues, considerable attempts have been made to define the problem by proposing terminologies and definitions. Keywords such as "transparency," "interpretability," "explainability", "intelligibility", "(white or grey or black)-Box", "responsible-AI", "third-wave-AI", "comprehensibility", are related to the main concept and have been used interchangeably [46], [45], [47], [48]. However, it is found that there are improper uses of the terms [46] (or perhaps "terminology ambiguity" [47]) in literature, and there is a lack of consensus on the definitions of the concept among researchers [45]. The author in [49] argues that interpretability is not a "monolithic concept" and has many ideas. Two notions of interpretability were proposed, such as transparency (i.e., how does the model work?) and post-hoc interpretability (i.e., what else can the model tell me?). Transparency consists of three main properties, such as Simulatability, Decomposability, and Algorithmic Transparency. Post-hoc interpretability presents techniques such as text explanations, visualization, and local explanations. In a recent work [46], the authors define "explainability" as "given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand". Other definitions can be found or compiled in [46], [45], [48].

In this light, the need for various forms of explainability has triggered the introduction of various regulations, guidelines, and standards by policymakers [26], [25], [50]. For instance, the "Right to Explanation" [51], [52] legal demands in The European Union General Data Protection Regulation (GDPR) [26], in which it provides individuals the rights to obtain "meaningful information about the logic involved" in fully automated decisions [52]. This, in turn, has led to organizations being told to ready themselves in order to be able to meet such new requirements. According to the Senior Managing Director at Protiviti (a global internal audit and risk consulting firm), Shaheen Dil, "The increased complexity of machine learning models can create unique challenges for validation teams. Validators need to be prepared to use alternative methods or develop custom methods to meet regulatory requirements" [53].

To address the explainability concerns and to comply with legal demands, various strategies have been proposed. These approaches can fall under two main categories: intrinsically interpretable methods and post-hoc interpretability methods. In the former, the aim is to develop methods that are transparent by design. For instance, classical methods, such as Decision Trees and K-Nearest Neighbours, demonstrate higher levels of transparency [46]. In the latter, the target is to enhance the interpretability of the methods using other techniques (i.e., independent algorithms) to generate explanations, such as text explanations and visualizations. For instance, applying LIME [54], a well-known approach to estimate the black-box model predictions. Yet, one might ask, can we (as users) faithfully rely on these algorithmic aids for the explanations? Doesn't this require the explanations themselves to be explained in order to gain the skeptical user's trust?

It is important to note that there is no general consensus on how to evaluate and measure the interpretability in machine learning. Researchers with different explainability targets apply different assessment metrics and measures [47]. For instance, in [55], the authors developed a qualitative method based on transparency properties in [49] to assess the interpretability of a given approach.

V. Literature Survey of Topic Detection Approaches in the Context of Transparency and Carbon Footprint

In order to examine the existence of any Transparency or Carbon Footprint assessments in the published topic detection studies, the following methodology was applied:

1) Aggregate all topic detection approaches that were reviewed in the summarised surveys in Section II. A dataset of 64 approaches in total was gathered. Readers interested in the various taxonomies of the topic detection approaches are referred to [29], [1], [30], [13], [14].
2) For each approach, examine the existence of any assessment related to the two qualities by answering the following two questions (Yes/No):
   - Is there any Transparency related assessment?
   - Is there any Carbon Footprint related assessment?

None of the surveyed approaches presented any form of Transparency or Carbon Footprint assessments, except the work in [28], where the Performance and Transparency of the proposed and benchmarked methods were evaluated.

Thus, due to justifications that are presented in section III and IV, and to bridge the gaps that are shown in section II and V, the rest of this paper provides an experimental analysis on the prominent topic detection approaches based on the three qualities of Performance, Transparency, and Carbon Footprint.

VI. Experiments

This section starts by introducing the general experimental setup (including: environment, corpuses, pre-processing procedure, and topic detection approaches) for the three assessments (i.e., Performance, Carbon Footprint, and Transparency). Then, it presents the evaluation methodology and the results of each of them.

A. Experiment Environment

All codes for the experiments were developed by using C#, Python, and Structured Query Language (SQL) programming languages. For details about the hardware and software configuration, see Table I.

TABLE I: Hardware and Software Configurations

Platform system	Windows-10-10.0.19043.1654
Installed RAM	64.0 GB (63.8 usable)
GPU model	NVIDIA GeForce GTX 1650 Ti
Processor	Intel(R) Core(TM) i9-10885H CPU @ 2.40GHz 2.40 GHz
Hard Disk	954 GB (601 GB free)
Software	Visual Studio 2019 (community), SQL Server Express 2017, SQL Management Studio (V18.2)

B. Experiment Dataset

Experiments were conducted using two different datasets: Text REtrieval Conference (TREC 2015 Microblog Track), available at [56], and EVENT2012, available at [57]. The TREC 2015 Microblog Track corpus comprises of about 6,191 thousand annotated tweets distributed across 46 topics. The EVENT2012 dataset, which was created in 2012 by the efforts in [58], contains 120 million tweets. 506 events are manually linked to more than 150 thousand tweets. We managed to download 3,195 thousand tweets (out of 6,191 thousand) and 69,553 thousand tweets (out of 101,239 thousand) through Twitter API, for TREC dataset and EVENT2012 dataset, respectively. Various reasons could be behind this, for instance, a tweet could be removed by the user who posted them, or a Twitter's account could be set to private.

C. Experiment Details

1) Text pre-processing: Each tweet in the two datasets underwent the same (simple) pre-processing procedure, summarised as follows:

- Remove hyperlinks and any non-alphabetic or non-numeric characters (except "#").
- Change all characters to lower case letters.
- Tokenize texts based on the white space between any set of characters.
- Remove stop words, this is conducted using Microsoft.ML library [59].

2) Topic detection methods :

a) Contextual Analysis Approach (CA): The contextual analysis (CA), introduced in [60], is an unsupervised approach that builds a hierarchical structure (called Hierarchical Knowledge Tree) to capture words' relationships (based on their contextual appearance) for a given collection of texts (documents). The constructed tree comprises of two main types of containers: a node and a HKT. The node groups the words that co-exist in a similar set of documents (governed by a user-defined threshold). Every node is located in the HKT container, which encapsulates at least one node. This container accommodates the nodes that contain words with similar strength (based on their occurrence in the given corpus). To form a parent-child contextual relationship, all documents that are used to construct a given node are used to construct its Child-HKT and Child-Node(s) (for more details about this approach the reader is referred to the original work [60] and [61]).

In recent work, Al Sulaimani and Starkey [28] presented an approach based on CA for a supervised short text classification problem. According to the assessments, the method shows its competitive capabilities for the task, and the transparency assessments reveal that the method is simple and transparent.

CA creates the tree-like structure in an unsupervised manner, however, up to date no assessment has been conducted to assess its capabilities for the topic detection problem without using the labeled dataset. Therefore, due to: (1) the potential capabilities presented in [28], (2) the transparency aspects of CA, this method is selected among the other prominent approaches for the experiments in this paper.

To examine the potential topic detection capabilities of CA, the following assumptions (definitions) are adopted (see Fig. 2). Two main types of containers are shown: HKT and Nodes. While, the HKT container encapsulates one or more nodes, the node container accommodates words and their sources. There are three distinct types of HKT: Seed-HKT, Child-HKT, and Refuge-HKT. Moreover, Seed-Node, Child-Node, Refuge-Node, and Orphan-Node are the four main node types.):

- **Seed-Node**: is a container that encapsulates one or more words as well as the documents (tweets) they appear in. This node is located in the Seed-HKT and it represents a main topic in the corpus. A tree must have at least one Seed-Node.
- **Seed-HKT**: is a container that highlights the most important words in related documents (tweets) in a corpus that belong to particular topics or categories. These main topics are represented as Seed-Nodes in a tree. A tree must have one Seed-HKT.
- **Child-Node**: is a container that encapsulates one or more words as well as the documents (tweets) they appear in. This node is located in the Child-HKT and it represents the sub-topics of the parent topic. A tree can have one or more Child-Nodes.
- **Child-HKT**: is a container that highlights other important words in the related documents (tweets) that formed the parent's topic or category and belong to particular sub-topics or sub-categories of the parent topic or category. These sub-topics are represented as Child-Nodes in the Child-HKT. A tree can have one or more Child-HKTs, each must be linked to one parent node either a Seed-Node or a Child-Node.
- **Refuge-Node**: is a container that encapsulates documents (tweets) that are in the corpus but none of their words appear in their Sibling-Nodes. These documents (tweets) cannot form a topic or a category similar to the strength of their Sibling-Nodes. Any HKT can have at most one Refuge-Node.
- **Refuge-HKT**: is a container that highlights other important words in the related documents (tweets) in a corpus that belong to particular sub-topics or sub-categories of the parent's topic or category. These topics are represented as Child-Nodes in the Refuge-HKT. A tree can have one or more Refuge-HKTs, each must be linked to one Refuge-Node.
- **Orphan-Node**: is a container that encapsulates one or more words and the documents (tweets) they appear

in. This node is located in a Refuge-HKT in which its parent is a Refuge-Node and none of its ancestors is a Child-Node or a Seed-Node.
- **Path:** is any node sequence from a starting node to any of its descendants Child-HKT or any specific descendants Child-Node in the tree along the parent-child connections. It must contain at least one node. The path represents a link between the topics and their sub-topics.

For the sake of the experiments in this paper, and to fairly compare with the other methods, we considered the nodes that appear in the Seed-Nodes and the Orphan-Nodes to be the generated clusters of the corpus.

b) K-Means: K-Means algorithm is one of the most widely implemented methods for clustering problems [62]. It partitions a dataset into (K) different clusters (C) based on the distance of each data point (x) from the closest cluster center (μ), or centroid, where each cluster is disjoint all other clusters. To accomplish this task, the number of clusters (K) needs to be specified in advance as well as the appropriate distance function should be selected, such as, Euclidean distance, the dot product similarity or cosine similarity. However, the standard K-Means goal is to minimize the sum of squared distances over all clusters:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - u_k\|^2 \quad (1)$$

The main steps of this algorithm are as the following:

1) Initialise cluster centres.
2) Assign data points (N) to the closest cluster center.
3) Recompute all cluster centers as the mean of assigned data points.
4) Repeat 2 and 3, (I) number of times.

K-Means is a simple and an efficient approach that has been empirically examined to solve various clustering problems [63]. Yang and Rayz [64] stated that this algorithm was a scalable method when it was implemented to solve event detection problem from a Twitter dataset based on the occurrence of hashtags. Also, Lu and Zhou [65] managed to implement K-Means to analyze Twitter data in order to predict evolution of the hurricane Sandy. The author analyzed 74,000 tweets and clustered them based on their location and time details.

It is noteworthy, as it has been stated above, that the first step of the K-Means algorithm is to indicate the number of initial clusters (K). Unfortunately, there is no mathematical process to guide this selection [63]. In practice, this can be a challenging task because a random selection of clusters can lead to a different clustering outcome. Also, the result of K-Means algorithm can be worsened if the data points include outliers (i.e., data points that differ significantly from other data points), because this can lead to poor clustering. On the other hand, there are different approaches that can help to enhance the initialization step, for instance, K-Means++ algorithm proposed by the work in [66].

c) Latent Dirichlet Allocation (LDA): Latent Dirichlet Allocation (LDA), introduced by Blei *et al*. [67], is a commonly used algorithm to automatically capture latent topics in a documents corpus. It is a probabilistic approach, based on bag of words, to build topic-per-document and word-per-topic models, in which a document is a probability distribution over topics and a topic is a probability distribution over words.

LDA is based on the assumption that a document is created on what is called, a generative process [68], which includes two main steps: (a) choosing topic proportions over fixed topics; (b) generating words based on the topic mixtures and the corresponding distribution over words.

More formally, the LDA model can be described as:

1) For each topic (k) in [1,...,K], sample β_k from Dirichlet(η).
2) For each document (d) in [1,...,D], sample θ_d from Dirichlet(α).
3) For each word w_{dn} , where d $\in$ [1,...,D] and n $\in$ [1,...,N]
4) choose a hidden topic z_{dn} from Multinomial (θ_d).
5) choose a word w_{dn} from Multinomial ($\beta_{z_{dn}}$).

For more details, readers are referred to the original work [67] and [68].

d) Nonnegative Matrix Factorization (NMF): Nonnegative Matrix Factorization (NMF) [69] is a well-known approach to decompose a non-negative matrix into (approximated) two lower dimensional non-negative matrices. For a given matrix $H \in \mathbb{R}^{m \times n}$, it computes an approximate factorisation such that:

$$H \approx UV \quad (2)$$

where $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{k \times n}$ are two factors, with the following optimisation problem [70]:

$$Min_{U,V} \left\| H - UV^T \right\|_F^2 \quad (3)$$

For the topic modeling task, NMF receives a document-term matrix in order to find a document-topic and a topic-term matrices. The number of k latent topics is given to the algorithm.

e) BERTopic: More recently, the author in [71] proposed a topic modelling approach, called BERTopic. It is based on the text embedding techniques, i.e., BERT-based, to overcome the limitation of other prominent topic modeling methods (such as LDA and NMF), in which the context of words is not well considered.

The main steps of BERTopic, are:

1) By using pre-trained language models, generate embeddings for the underlying corpus.
2) Reduce the dimensions of embeddings by implementing UMAP [72] technique.
3) Cluster the documents' representations using HDBSCAN [73] algorithm.
4) Extract topics from the generated clusters by using a modified version of Term Frequency - Inverse Document Frequency (TF-IDF) approach (which emphasizes the words that appear frequently in a document and appear rarely in a corpus). From each cluster, a single document, that compiles all encapsulated documents, is generated. Then, a class-based TF-IDF [71] is computed according to 4.

$$W_{t,c} = TF_{t,c} \log\left(1 + \frac{A}{TF_t}\right) \quad (4)$$

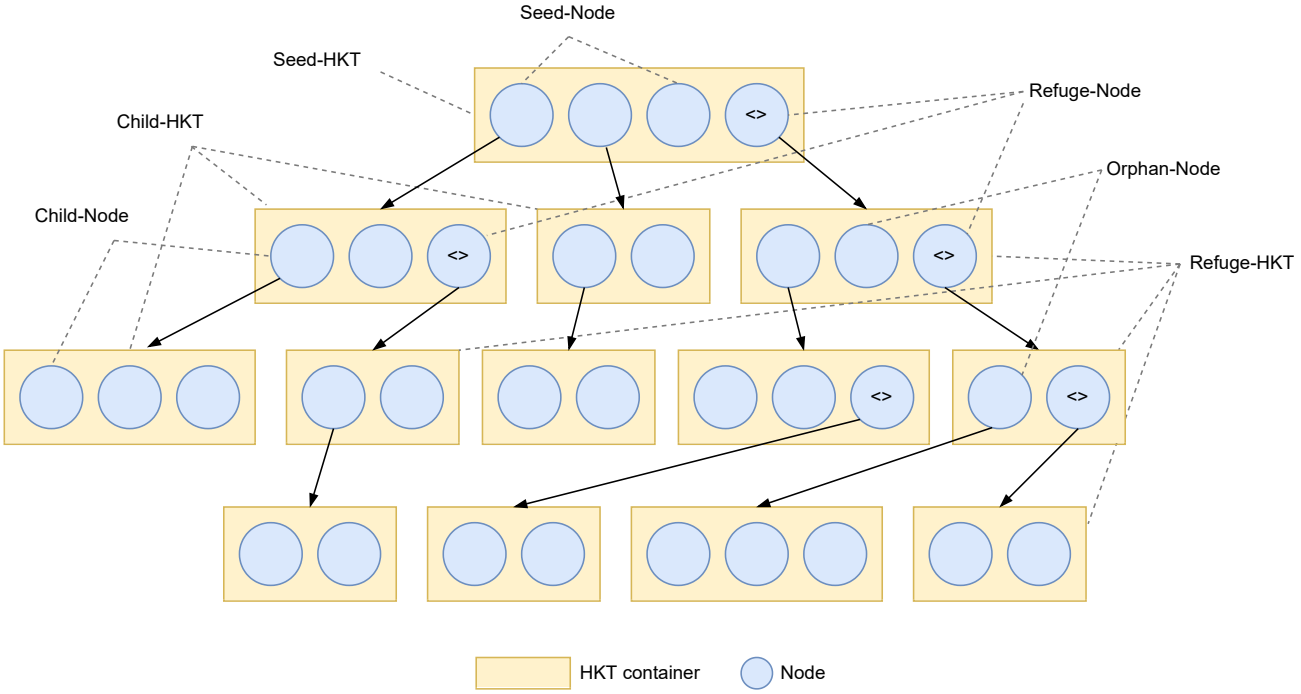

Figure 2. Contextual Analysis (CA) components.

where $TF_{t,c}$ is the frequency of term t in class c. A is the average number of words per class. TF_t is the frequency of term t in all classes.

Note: for the experiments in this paper, the implementation by [74] is used for K-Means[2], Latent Dirichlet Allocation (LDA)[3], Nonnegative Matrix Factorization (NMF)[4]. For BERTopic, we used the code in [75]. For Contextual Analysis (CA) approach, we implemented the algorithm (in C#) that is described in the original work [60].

D. Experimental Design

In order to assess the Performance, the Carbon Footprint, and the Transparency of the five approaches, three different experiments were designed. The following subsections provide the details of each of them.

To conduct a comparative analysis of the methods, two dataset themes with distinctive characteristics were adopted. Each theme was designed with different settings. For TREC 2015 (Theme1), we created ten sub-datasets from the main corpus for the period from 20-July-2015 to 29-July-2015. Each comprises of tweets of a single day (i.e., one sub-dataset per day). For each day, any event that contains less than three tweets was removed. For Event2012 (Theme2), with the similar approach, 28 different sub-datasets from the main corpus were selected. Each dataset contains tweets for a single day, for the period from 10-October-2012 to 6-November-2012. Fig. 3 gives a summary of the two themes and their different characteristics. While chart (a) shows the number of words found in each tweet, chart (b), chart (c), and chart (d) present (per experiment) the number of tweets, the number of events, and the imbalance ratio of the events, respectively. The imbalanced ratio was computed by dividing the number of samples (tweets) in the majority class (event) over the number of samples in the minority class. On average, there are 954 tweets per experiment and 9 words per tweet in Theme1, and there are 2500 tweets per experiment and 8 words per tweet in Theme2. (Note: these charts display the details after the text pre-processing phase)

E. Evaluation Methodology

1) Performance assessment: To evaluate and compare the results of the methods, the metrics that were used in [30] were applied, namely, purity, normalized mutual information (NMI), and BCubed f1 (see Eqs. (5), (6) and (12)). These measures evaluate the quality of the produced clusters utilizing the labels in the annotated dataset.

Purity: By labeling each produced cluster by the dominant category (i.e., the most common category in the encapsulated tweets), the quality is scored as:

$$Purity(W,C) = \frac{1}{N}\sum_{k=1}^{K}\max_{j}|w_k \cap c_j| \quad (5)$$

where W is the set of clusters, C is the set of categories, w_k: the set of tweets in the category k (i.e., according to the annotated label), c_j: the set of tweets in the cluster j and N is the number of tweets in the dataset.

NMI: To compute the tweets' clusters quality using NMI, the following formula is used:

$$NMI = \frac{2I(W;C)}{H(W)+H(C)} \quad (6)$$

where I(W;C) is the mutual information between W and C (see Eq. (7)), H(W) is the average entropy of the categories

[2]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[3]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html.

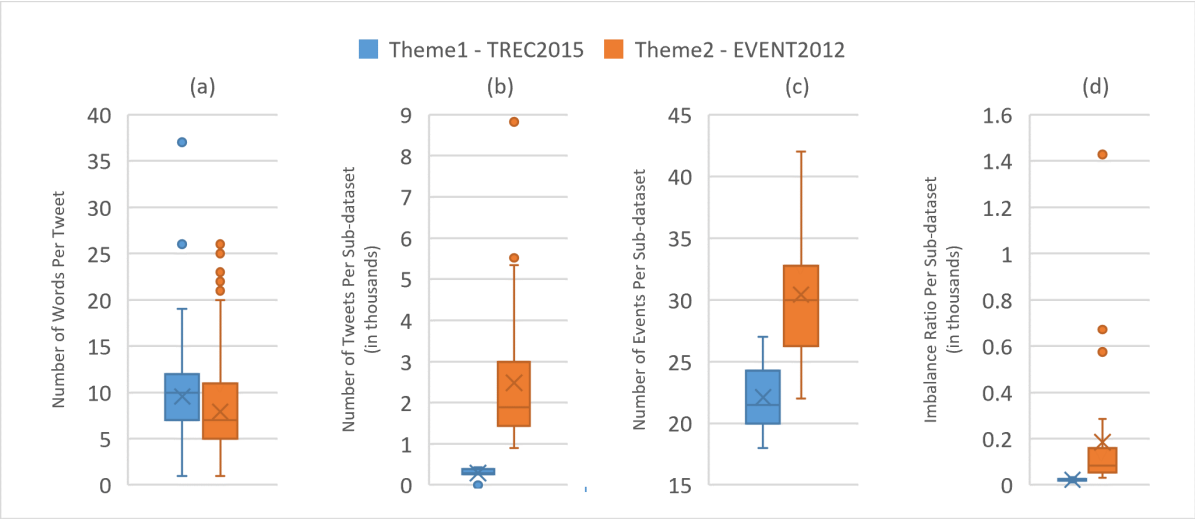

Figure 3. Summary of dataset setup. A total of ten sub-datasets (Theme1) and 28 sub-datasets (Theme2) were created using TREC2015 and EVENT2012 corpuses, respectively.

(see Eq. (8)), and H(C) is the average entropy of the clusters(see Eq.(9)).

$$I(W,C) = \sum_{k=1}^{K}\sum_{j=1}^{J} \frac{|w_k \cap c_j|}{N} \log \frac{N|w_k \cap c_j|}{|w_k||c_j|} \quad (7)$$

$$H(W) = -\sum_{k=1}^{K} \frac{|w_k|}{N} \log \frac{|w_k|}{N} \quad (8)$$

$$H(C) = -\sum_{j=1}^{J} \frac{|c_j|}{N} \log \frac{|c_j|}{N} \quad (9)$$

BCubed f1: The BCubed f1 of any tweet in the dataset is the average BCubed precision and BCubed recall (see Eqs. (10), (11), (12)). The BCubed precision for a tweet captures the number of tweets in the cluster that have the same category according to the annotated label. The BCubed recall measures the number of tweets in the same category found in the cluster. The overall f1 BCubed is the average f1 of all tweets in the dataset.

$$BCubedPrecision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (10)$$

$$BCubedRecall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (11)$$

$$BCubedf1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

For more details about these methods, readers are referred to the work in [76].

2) Carbon footprint assessment: To report the carbon dioxide equivalent (CO_2eq) emissions during the computational task of the five approaches, we adopt the CodeCarbon tool that is available in [77].

CodeCarbon is an open-source package that estimates the CO_2eq in kilograms using the following equations [43] :

$$CO_2eq = CarbonIntensity \times PowerUsage \quad (13)$$

where $CarbonIntensity$ and $PowerUsage$ are the consumed electricity (in kgCO_2/kWh) and power (in kWh), respectively, for computation.

A combination of energy sources, including fossil fuels (like natural gas) and renewables (like solar), determines the carbon intensity. One of three techniques is used to estimate this value:

- The carbon intensity of electricity per cloud provider or country is utilized, if available.
- Alternatively, intensities are calculated by using predefined values per energy source (e.g., Petroleum = 816 kg/MWh, Wind = 26 kg/MWh) and their proportional usage. This computation is performed using the following equation::

$$NetCarbonIntensity = \sum_{AllEnergySources} CarbonIntensity_{Source} \times Percentage_{Source} \quad (14)$$

- In some cases, a world average value is considered.

For each method, we run the tool to track the estimated released emissions during the training process of the Theme2-EVENT2012 sub-datasets (see VI-D). We repeated the experiments 10 times, and the minimum emission value of each method was recorded. Note: the experiments were conducted in Scotland, United Kingdom.

3) Transparency assessment: To assess the transparency of the five methods, we adopted the approach that is proposed in [55] and used in our previous work [28]. Based on the proposed properties of transparency in [49], namely "simulatability", "decomposability", and "algorithmic transparency", the authors assess the methods according to the following three questions:

- For simulatability: "Is the entire model simple enough to be fully understood by a user?"
- For decomposability: "Is each part of the model (each input, parameter, and calculation) intuitively explainable?"

- For algorithmic transparency: "Is the algorithm deterministic (non-stochastic) without using any random numbers?"

Similar to [28], we restricted the sample space for assessing the simulatability to 10 observations (i.e., tweets).

The definition suggested in [46] is that the requirements of the target audience should be considered as a key concept to meet the explainability (see section IV). Note that we consider the terms explainability and interpretability in this context to be equivalent. The decisions reached by methods should meet the level of understanding that is required by the intended audience, and therefore in a human-readable format. However, we feel that these definitions require further refinement. In terms of explainability, this is linked explicitly to the question of the data that is being asked. For example, if the question is to ask what are the key words in the topic, then all of the methods are capable of returning a list of words which means that they are capable of explaining how they identify that topic. However, if the question is extended beyond this narrow definition to ask what the key words are and what is their relationship to each other, then they are not capable of answering this question - beyond a simple statistical measure of their strength for that particular topic.

Therefore the issue of explainability is explicitly linked to the requirement of the user and what particular question that they are asking of the data. The definition of the user question, and whether the method is capable of answering it is, therefore, a binary output of: yes it can; or no it cannot.

F. Results & Discussion

1) Performance assessment: The performance of the two themes experiments (i.e., Theme1 using the TREC2015 Microblog Track dataset, and Theme2 using the EVENT2012 dataset) are summarised in Fig. 4. The average Purity, NMI, and BCubed f1 scores of the experiments are shown.

In Theme1 (see part Theme1-TRECT2015 in Fig. 4), on average, the BERTopic approach outperforms the other four methods by achieving the highest scores in the three metrics 92.3%, 89.1%, and 81.3% in purity, NMI and BCubed f1 assessments, respectively. This is about 11% and 6% higher than the second best performing algorithm, i.e., Contextual Analysis, in NMI and BCubed, respectively, with very similar results in purity. NMF produces competitive NMI scores with Contextual Analysis, however, its purity and BCubed measures are lower by about 12% and 20%, respectively. LDA achieves the worst scores.

In Theme2, Contextual Analysis is better than all the others, with 89.1%, 80.4%, and 70% in purity, NMI, and BCubed, respectively (as shown in Fig. 4 in part Theme1-EVENT2012). Interestingly, K-Means, LDA, and NMF show remarkable improvements in this theme. However, an opposite trend for BERTopic was observed.

2) Carbon footprint assessment: The Carbon Footprint assessment's results are summarized in Table II (the accumulative results are shown). BERTopic is the most carbon-intensive approach. Contextual Analysis, LDA and NMF produced very similar emissions, which are lower by 0.39-0.44 grams than the produced emissions by K-Means. It is important to note that the estimated BERTopic's CO_2eq does not include the contributed emissions during the training process of BERT model, which will be considerable.

TABLE II: Results of Emissions Released by Five Methods during Training Tasks

Method	Duration (in second)	CO_2eq Emissions (in gram)
Contextual Analysis	130	0.492
K-Means	243	0.909
Latent Dirichlet Allocation	130	0.519
Nonnegative Matrix Factorization	126	0.469
BERTopic	741	2.56

Although the values reported in this example are small compared to those of other industrial activities, the cumulative effects of these emissions and their potential environmental harm should be considered, especially when integrated into various daily life or large-scale applications. With the rapid development of such techniques, careful attention should be paid to optimizing resource utilization, reducing the carbon footprint associated with energy-intensive computations, and showing commitment to environmental sustainability efforts.

3) Transparency assessment: Table III shows the results of the transparency assessments. None of the methods, except Contextual Analysis, satisfies the three qualitative assessments, proposed in [55]. Contextual Analysis algorithm is considered transparent in all components. The approach does not require any technical background to comprehend the generation process of the tree. Although the main steps of K-Means algorithm are simple, the underlying objective to minimize the sum of the squared distances between data points and centroids requires some other tools to explain. Understanding the modeling process by LDA and NMF requires mathematical and statistical background. The four components of BERTopic, i.e., HDBSCAN, BERT, UMAP, and TFIDF, make it the most complicated approach among the others.

It is important to highlight that other researchers focused on the interpretability of the outputs from the methods. The authors in [78] automatically measured topic interpretability of LDA's results (with various pooling[5] strategies using Twitter) based on human judgments on the semantics of words. They found that the interpretability can be significantly improved without changing LDA internal mechanisms, especially when using a Hashtag-based pooling approach. The efforts in [22] compared the topics that are provided by NMF and K-Means, using world cup tweets. By visualizing the most frequent words, they found that NMF offers "more easily interoperated results". This agrees with the conclusion drawn by [23], in which the results of four algorithms (NMF, K-Means, KMedoids, and DBSCAN) were investigated. However, in [79], it was reported that LDA produced more interpretable topics than NMF, using manual human inspection of the generated results. All of these previous studies measure interpretability against the ability of the method to return a list of important words only.

[5]Pooling here means that tweets are grouped as a single document to be fed to the LDA method, for instance, group tweets by Hashtag.

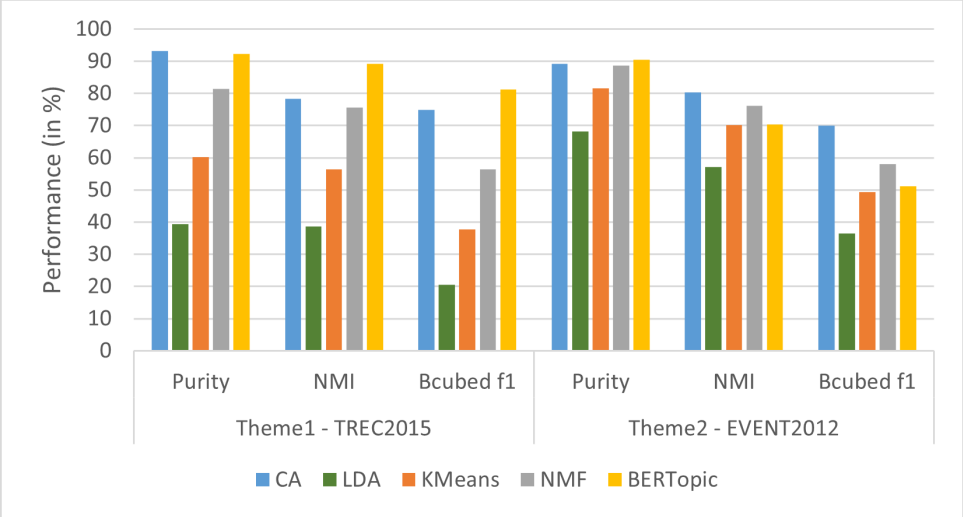

Figure 4. Performance comparison between five approaches on Theme1 and Theme2.

TABLE III: Transparency Assessment's Results of the Five Approaches

Algorithm	Simulatability: "Is the entire model simple enough to be fully understood by a user?"	Decomposability: "Is each part of the model (input, parameter, and calculation) intuitively explainable?"	Algorithmic Transparency: "Is the algorithm deterministic (non-stochastic) without using any random numbers?"	Potential for human-readable output on decision making process?
Contextual Analysis (CA)	✓A human can follow tree creation steps in a reasonable time.	✓Variables are preserved during the tree generation process, and the required parameters and the calculations involved do not require technical skills.	✓deterministic	✓
K-Means	✗Although the algorithmic steps and similarity measure under use are simple to comprehend, the repetition of steps is beyond the non-technical user's capabilities to simulate.	✗The objective computation to minimize the sum of the squared distances between data points and centroids is difficult to understand by a non-technical user.	✗non-deterministic	✓
Latent Dirichlet Allocation (LDA)	✗Statistical relationships and the inference procedure can not be understood by a non-technical user.	✗can not be understood by a non-technical user.	✓deterministic	✗
Nonnegative Matrix Factorization (NMF)	✗Decomposing the matrix requires a mathematical background.	✗Understanding the process of finding the approximate decomposition requires mathematical skills.	✗non-deterministic	✗
BERTopic	✗The dependency on four different approaches (such as: HDBSCAN, UMAP, TF-IDF, and BERT) that require a technical and mathematical background.	✗The underlying approaches and their internal interactions are difficult to understand.	✗non-deterministic	✗

In this regard, we further investigated the generated topics by each approach. LDA, NMF, K-Means, and BERTopic capture the words for each topic in a flat representation. For example, Fig. 5 presents a comparison of the top ten words that were generated for six different events by each method, using the subdataset Theme2-EVENT2012-(19-October-2012). The node with the symbol "<>" represents data not matching the words shown at the same level, i.e., a refuge node. As illustrated in the figure, "timberlake", "justin", "biel", "married", "eonline", are the common words (among the top ten most important words) that were produced by the five methods for Event 9, which is an event about the marriage of Justin Timberlake (an American singer and actor) and Jessica Biel (an American actress and model), as described in the original corpus. Note: the best matching topic for each event was selected manually, and the words are ordered (from left to right) according to their importance which is demonstrated in each method's output. This shows that these methods are similar in that they can identify a number of key words that occur in topics together, but does not give any further information regarding their relative importance. So, in other words, the methods can be considered explainable when they answer the question: what words are important for the topic? On the other hand, the methods are not considered explainable (since they cannot give an answer) for the question: what is the relationship between the words that are important for a topic? However, Contextual Analysis differs in the

way it represents the words relationship to a topic. By its intrinsic design (and without using any other methods), it gives a hierarchical tree structure of the results. This representation simply highlights the important concepts at the upper layers in a tree-like structure. More granular details of any given concept can be revealed by navigating through its lower level concepts. Also, the words that are found in a similar set of sources are grouped in a single node. To illustrate, the words "justin" and "timberlake" are grouped in one node at a higher level, and the words "jessica" and "biel" are encapsulated in its Child-Node. The appearance of Justin Timberlake at a higher level of the tree representation of the event may give various insights. It could be related to the level of fame achieved by him, compared to Jessica Biel, at that time, or other related facts that can be linked to the event. Therefore CA is explainable for both questions (What words are important for the topic? What is the relationship between the words that are important for a topic?) since it can return an answer that can be understood by a human.

VII. Conclusion and Future Work

In this review, we provided an overview of two emerging fields in the realm of Artificial Intelligence, i.e., "Right to Explanation" and "Green AI". We justified the need for environmental-friendly algorithms and the importance of their Transparency aspects. Also, we highlighted the international momentum toward fulfilling the new regulatory requirements in these regards.

Then, we conducted a survey on the topic detection approaches to explore the availability of any Transparency or Carbon Footprint assessments in the previous works. Up to now, we could not find any review that is focused on these two important topics. Also, none of the previous works in the topic detection assessed the proposed methods based on these two qualities, except the work in [28], in which the Transparency assessment only was conducted.

Based on the three qualities: Performance, Carbon Footprint, and Transparency, the experimental work focused on assessing five methods for the topic detection task, including: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), K-Means, BERTopic and Contextual Analysis (CA). Two sub-datasets themes (with different characteristics), using two corpuses, were carefully designed. The results of the conducted experiments on Theme1 show that BERTopic is the best-performing approach overall, however, in Theme2 Contextual Analysis method achieves the best scores. For the Transparency assessment, BERTopic, like three other methods, namely: LDA, NMF and K-Means, fails to satisfy the Transparency assessment. Also, BERTopic is the most carbon-intensive approach even without consideration of the energy required for the training of the BERT model itself. Overall, Contextual Analysis (CA) is the only method that satisfies the checklist, with a very competitive performance, towards a transparent and environmental-friendly approach for short text topic detection task.

Although the Contextual Analysis approach gives rich information about a certain topic and has been shown to be explainable for the two questions (What words are important for the topic? What is the relationship between the words that are important for a topic?) it is not clear how this method can be utilized to study how each topic evolves over time, or whether the ability to answer the second question of the relative importance of words is particularly useful. The current algorithm builds a tree-like structure for the provided dataset (whether for every hour or day or week), in a one-go procedure, to capture the relationship between the words based on their appearance in the same context. Details about the progression of any topic within a timeframe are lost. Preserving and representing this valuable information can offer important insights about a topic. Thus, our future work will focus on how the Contextual Analysis algorithm can be utilized or modified (by changing its basic machinery and without harming its level of transparency), in order to provide information about the evolution of topics over time.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Sami Al Sulaimani conducted the research, created, tested the experiments, and drafted the paper. Andrew Starkey reviewed, prepared/edited the manuscript and supervised the research. All authors had approved the final version.

References

[1] Z. Saeed, R. A. Abbasi, O. Maqbool, A. Sadaf, I. Razzak, A. Daud, N. R. Aljohani, and G. Xu, "What's Happening Around the World? A Survey and Framework on Event Detection Techniques on Twitter," *J. Grid Comput.*, vol. 17, no. 2, pp. 279–312, Jun 2019. [Online]. Available: https://doi.org/10.1007/s10723-019-09482-2

[2] We Are Social Ltd, "DIGITAL 2022: ANOTHER YEAR OF BUMPER GROWTH." [Online]. Available: https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/ (Accessed 2022-05-16).

[3] M. Avvenuti, S. Cresci, F. D. Vigna, and M. Tesconi, "On the need of opening up crowdsourced emergency management systems," *AI Soc.*, vol. 33, pp. 55–60, Feb 2018. [Online]. Available: https://doi.org/10.1007/s00146-017-0709-4

[4] C. Zong, R. Xia, and J. Zhang, "Topic Detection and Tracking," in *Text Data Mining*. Singapore: Springer Singapore, 2021, pp. 201–225. ISBN 978-981-16-0100-2. [Online]. Available: https://doi.org/10.1007/978-981-16-0100-2_9

[5] Twitter Inc., "Counting characters." [Online]. Available: https://developer.twitter.com/en/docs/counting-characters (Accessed 2022-06-24).

[6] K. Byrd, A. Mansurov, and O. Baysal, "Mining Twitter data for influenza detection and surveillance," in *Proc. 2016 IEEE/ACM Int. Work. Softw. Eng. Healthc. Syst.* Austin, TX, USA: ACM, May 2016. ISBN 9781450341684 pp. 43–49. [Online]. Available: https://doi.org/10.1145/2897683.2897693

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *Proc. 19th Int. Conf. World Wide Web*, ser. WWW '10. Raleigh, NC, USA: ACM, 2010. ISBN 978-1-60558-799-8 pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777

[8] N. Alsaedi, P. Burnap, and O. Rana, "Can We Predict a Riot? Disruptive Event Detection Using Twitter," *ACM Trans*. Internet Technol., vol. 17, no. 2, pp. 1–26, May 2017. [Online]. Available: https://doi.org/10.1145/2996183

[9] M. Wang and M. S. Gerber, "Using Twitter for Next-Place Prediction, with an Application to Crime Prediction," in *2015 IEEE Symp. Ser. Comput. Intell.*, Cape Town, South Africa, 2015, pp. 941–948. [Online]. Available: https://doi.org/10.1109/SSCI.2015.138

[10] M. I. Mahmud, M. Mamun, and A. Abdelgawad, "A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning," in *2022 IEEE Glob. Conf. Artif. Intell.Internet Things*, Dec 2022, pp. 166–170. [Online]. Available: https://doi.org/10.1109/GCAIoT57150.2022.10019058

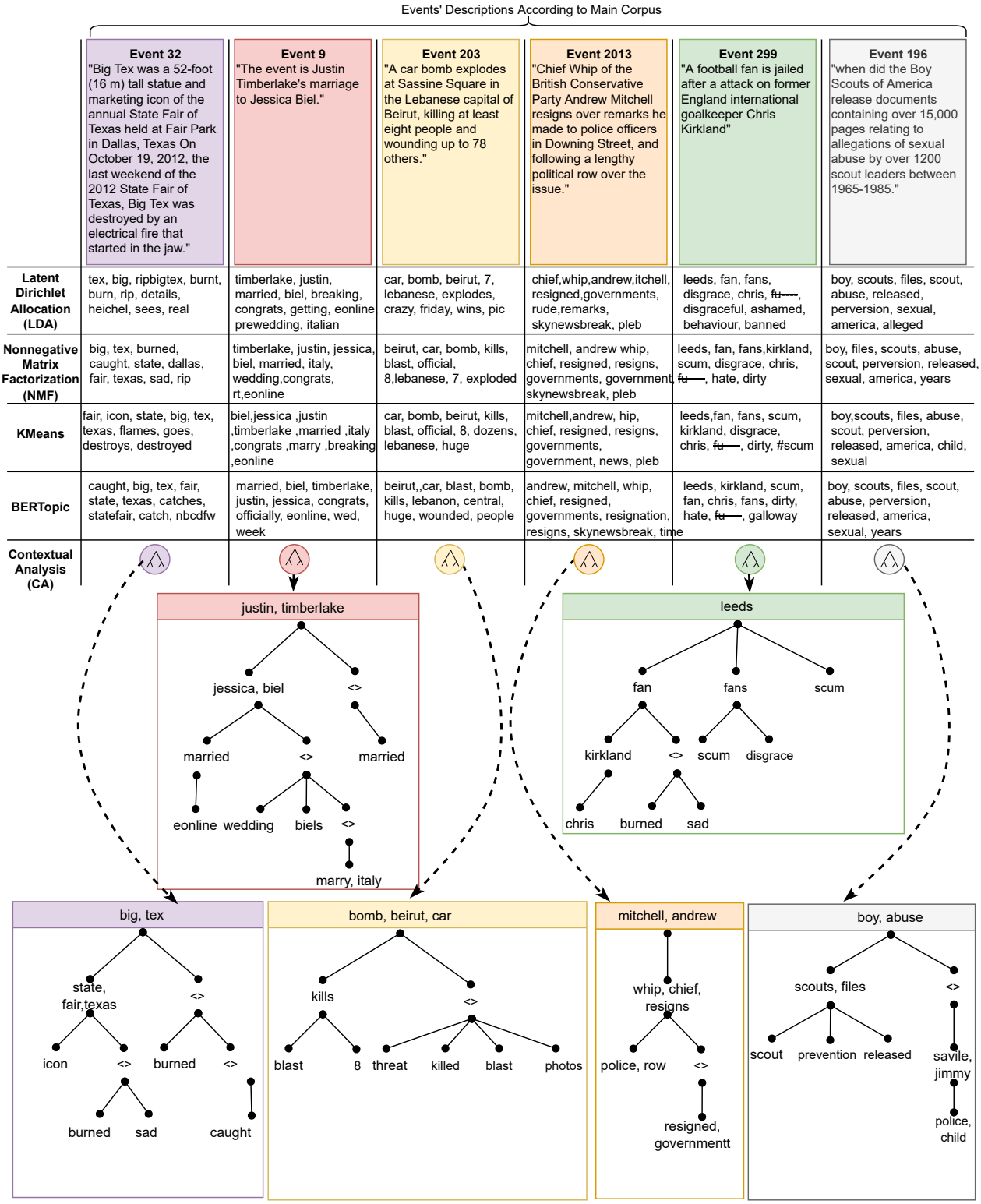

Figure 5. Comparison of the top ten words that were generated for six different events by each method.

[11] T. Pratama and A. Purwarianti, "Topic classification and clustering on Indonesian complaint tweets for bandung government using supervised and unsupervised learning," in *2017 Int. Conf. Adv. Informatics, Concepts, Theory, Appl.*, Aug 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICAICTA.2017.8090981

[12] V. K. Ayyadevara, "Word2vec," in *Pro Machine Learning Algorithms*. Berkeley, CA: Apress, 2018, pp. 167–178. ISBN 978-1-4842-3564-5. [Online]. Available: https://doi.org/10.1007/978-1-4842-3564-5 8

[13] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, "A review of approaches for topic detection in Twitter," *J. Exp. Theor. Artif. Intell.*, vol. 33, no. 5, pp. 747–773, 2021. [Online]. Available: https://doi.org/10.1080/0952813X.2020.1785019

[14] P. Tijare and J. Rani Prathuri, "A Survey on Event Detection and Prediction Online and Offline Models using Social Media Platforms," *Mater. Today Proc.*, 2021. [Online]. Available: https://doi.org/10.1016/j.matpr.2021.02.164

[15] S. Petrovi´c, M. Osborne, and V. Lavrenko, "Streaming First Story Detection with Application to Twitter," in *Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist.*, ser. HLT '10. Los Angeles, CA, USA: Association for Computational Linguistics, 2010. ISBN 1932432655 pp. 181–189.

[16] Q. Li, A. Nourbakhsh, S. Shah, and X. Liu, "Real-Time Novel Event Detection from Social Media," in *2017 IEEE 33rd Int. Conf. Data Eng.*, San Diego, CA, USA, Apr 2017, pp. 1129–1139. [Online]. Available: https://doi.org/10.1109/ICDE.2017.157

[17] M. Hasan, M. A. Orgun, and R. Schwitter, "Real-time event detection from the Twitter data stream using the TwitterNews+ Framework," *Inf. Process. Manag.*, vol. 56, no. 3, pp. 1146–1165, May 2019. [Online]. Available: https://doi.org/10.1016/j.ipm.2018.03.001

[18] D. Zhou, L. Chen, and Y. He, "An Unsupervised Framework of Exploring Events on Twitter: Filtering, Extraction and Categorization," *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, Feb 2015. [Online]. Available: https://doi.org/10.1609/aaai.v29i1.9526

[19] D. Metzler, C. Cai, and E. Hovy, "Structured Event Retrieval over Microblog Archives," in *Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.* Montréal, Canada: Association for Computational Linguistics, Jun 2012, pp. 646–655.[Online]. Available: https://aclanthology.org/N12-1083

[20] Q. Wang, J. Bhandal, S. Huang, and B. Luo, "Classification of Private Tweets Using Tweet Content," in *2017 IEEE 11th Int. Conf. Semant. Comput.*, San Diego, CA, USA, 2017, pp. 65–68. [Online]. Available: https://doi.org/10.1109/ICSC.2017.36

[21] E. Alabdulkreem, "Prediction of depressed Arab women using their tweets," *J. Decis. Syst.*, vol. 30, no. 2-3, pp. 102–117, Sep 2021. [Online]. Available: https://doi.org/10.1080/12460125.2020.1859745

[22] D. Godfrey, C. Johns, C. Meyer, S. Race, and C. Sadek, "A case study in text mining: Interpreting Twitter data from world cup tweets," *arXivPrepr. arXiv1408.5427*, 2014.

[23] M. Klinczak and C. Kaestner, "Comparison of Clustering Algorithms for the Identification of Topics on Twitter," *Lat. Am. J. Comput.*, vol. 3, no. 1, pp. 19–26, May 2016. [Online]. Available: https://lajc.epn.edu.ec/index.php/LAJC/article/view/99

[24] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist*. Florence, Italy: Association for Computational Linguistics, Jul 2019, pp. 3645–3650. [Online]. Available: https://doi.org/10.18653/v1/P19-1355

[25] E. Jillson, "Aiming for truth, fairness, and equity in your company's use of AI," 2021. [Online]. Available: https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai (Accessed 2022-06-24).

[26] Intersoft Consulting, "General Data Protection Regulation (GDPR)." [Online]. Available: https://gdpr-info.eu/ (Accessed 2022-06-20).

[27] The United Nations Framework Convention on Climate Change, "The Paris Agreement." [Online]. Available: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement (Accessed 2022-06-24).

[28] S. A. Sulaimani and A. Starkey, "Short Text Classification Using Contextual Analysis," *IEEE Access*, vol. 9, pp. 149 619–149 629, Nov 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3125768

[29] A. Weiler, M. Grossniklaus, and M. H. Scholl, "Survey and Experimental Analysis of Event Detection Techniques for Twitter," *Comput. J.*, vol. 60, no. 3, pp. 329–346, Mar 2017. [Online]. Available: https://doi.org/10.1093/comjnl/bxw056

[30] R. Nugroho, C. Paris, S. Nepal, J. Yang, and W. Zhao, "A survey of recent methods on deriving topics from Twitter: algorithm to evaluation," *Knowl. Inf. Syst.*, vol. 62, no. 7, pp. 2485–2519, Jul 2020.[Online]. Available: https://doi.org/10.1007/s10115-019-01429-z

[31] E. &. I. S. UK Government - Department for Business, "Sub-national electricity and gas consumption summary report 2020." [Online]. Available: https://www.gov.uk/government/statistics/sub-national-electricity-and-gas-consumption-summary-report-2020 (Accessed 2022-06-23).

[32] IPCC, Summary for Policymakers. in *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. Cambridge University Press, 2022, pp. 1–24. [Online]. Available: https://doi.org/10.1017/9781009157940.001

[33] International Energy Agency (IEA), "Net Zero by 2050." [Online]. Available: https://www.iea.org/reports/net-zero-by-2050 (Accessed 2022-06-24).

[34] The United Nations, "All About the NDCs." [Online]. Available: https://www.un.org/en/climatechange/all-about-ndcs (Accessed 2022-06-24).

[35] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," *arXiv Prepr. arXiv1910.09700*, 2019.

[36] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *J. Mach. Learn. Res.*, vol. 21, no. 248, pp. 1–43, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-312.html

[37] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," *AAAI*, vol. 34, no. 09, pp. 1393–13 696, Apr 2020. [Online]. Available: https://doi.org/10.1609/aaai.v34i09.7123

[38] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models," in *ICML Work. Challenges Deploying Monit. Mach. Learn. Syst.*, Jul 2020. [Online]. Available: http://arxiv.org/abs/2007.03051

[39] M. Yusuf, P. Surana, G. Gupta, and K. Ramesh, "Curb Your Carbon Emissions: Benchmarking Carbon Emissions in Machine Translation," *arXiv Prepr. arXiv2109.12584*, 2021.

[40] F. &. R. A. UK Government - Department for Environment and E. Agency, "Calculate the carbon dioxide equivalent quantity of an F gas." [Online]. Available: https://www.gov.uk/guidance/calculate-the-carbon-dioxide-equivalent-quantity-of-an-f-gas (Accessed 2022-06-23).

[41] U.S. Environmental Protection Agency, "Greenhouse Gas Equivalencies Calculator." [Online]. Available: https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator (Accessed 2022-05-01).

[42] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, and Others, "On the opportunities and risks of foundation models," *arXiv Prepr. arXiv2108.07258*, 2021.

[43] Mila, BCG GAMMA, Haverford College, and Comet, "CodeCarbon." [Online]. Available: https://codecarbon.io/ (Accessed 2022-04-15).

[44] L. Lannelongue, J. Grealey, and M. Inouye, "Green Algorithms: Quantifying the Carbon Footprint of Computation," *Adv. Sci.*, vol. 8, no. 12, p. 2100707, Jun 2021. [Online]. Available: https://doi.org/10.1002/advs.202100707

[45] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, Sep 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2870052

[46] A. Barredo Arrieta, N. D´ıaz-Rodr´ıguez, J. Del Ser, A. Bennetot,S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[47] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems, " *ACM Trans. Interact. Intell. Syst.*, vol. 11, no. 3–4, Dec 2021. [Online]. Available: https://doi.org/10.1145/3387166

[48] M.-A. Clinciu and H. Hastie, "A Survey of Explainable AI Terminology," in *Proc. 1st Work. Interact. Nat. Lang. Technol. Explain. Artif. Intell. (NL4XAI 2019)*. Association for Computational Linguistics, 2019, pp. 8–13. [Online]. Available: https://aclanthology.org/W19-8403

[49] Z. C. Lipton, "The Mythos of Model Interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, Sep 2018. [Online]. Available: https://doi.org/10.1145/3233231

[50] UK Government - Central Digital and Data Office, "Algorithmic Transparency Standard." [Online]. Available:

https://www.gov.uk/government/collections/algorithmic-transparency-standard (Accessed 2022-06-24).

[51] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017. [Online]. Available: https://doi.org/10.1609/aimag.v38i3.2741

[52] A. D. Selbst and J. Powles, "Meaningful information and the right to explanation," *Int. Data Priv. Law*, vol. 7, no. 4, pp. 233–242, 2017. [Online]. Available: https://doi.org/10.1093/idpl/ipx022

[53] Protiviti Inc., "Validation of Machine Learning Models: Challenges and Alternatives." [Online]. Available: https://www.protiviti.com/UK-en/insights/validation-machine-learning-models-challenges-and-alternatives (Accessed 2022-05-15).

[54] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, ser. KDD '16. San Francisco, CA, USA: Association for Computing Machinery, 2016. ISBN 9781450342322 pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[55] T. Mori and N. Uchihira, "Balancing the trade-off between accuracy and interpretability in software defect prediction," *Empir. Softw. Eng.*, vol. 24, no. 2, pp. 779–825, 2019. [Online]. Available: https://doi.org/10.1007/s10664-018-9638-1

[56] The National Institute of Standards and Technology (NIST) - U.S. Department of Commerce, "2015 Microblog Track." [Online]. Available: https://trec.nist.gov/data/microblog2015.html (Accessed 2022-01-01).

[57] University of Glasgow, "Twitter Event Detection Dataset." [Online]. Available: http://mir.dcs.gla.ac.u.k./resources/ (Accessed 2020-11-04).

[58] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a Large-Scale Corpus for Evaluating Event Detection on Twitter," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, ser. CIKM '13. San Francisco, CA, USA: Association for Computing Machinery, 2013. ISBN 9781450322638 pp.409–418. [Online]. Available: https://doi.org/10.1145/2505515.2505695

[59] Microsoft Corporation, "Microsoft.ML." [Online]. Available: https://www.nuget.org/packages/Microsoft.ML

[60] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," *IEEE Access*, vol. 8, pp. 17 722–17 733, Jan 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2958702

[61] A. A. Aziz, "Contextual-based approach for sentiment analysis," Ph.D. dissertation, Eng. School, Univ. Aberdeen, Aberdeen, U.K., 2020.

[62] C. Zhou and Q. Zhao, "Efficient Time Series Clustering and Its Application to Social Network Mining," *J. Intell. Syst.*, vol. 23, no. 2, pp. 213–229, 2014. [Online]. Available: https://doi.org/10.1515/jisys-2014-0005

[63] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun 2010. [Online]. Available: https://doi.org/10.1016/j.patrec.2009.09.011

[64] S.-F. Yang and J. Rayz, "An Event Detection Approach Based On Twitter Hashtags," *arXiv Prepr. arXiv1804.11243*, 2018.

[65] X. S. Lu and M. Zhou, "Analyzing the evolution of rare events via social media data and k-means clustering algorithm," in *2016 IEEE 13th Int. Conf. Networking, Sensing, Control*, Apr 2016, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICNSC.2016.7479041

[66] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," in *Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms*, vol. 8, New Orleans, LA, USA, Jan 2007, pp. 1027–1035.

[67] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[68] D. M. Blei, "Probabilistic Topic Models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr 2012. [Online]. Available: https://doi.org/10.1145/2133806.2133826

[69] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct 1999.

[70] C. C. Aggarwal, "Text Sequence Modeling and Deep Learning," in *Machine Learning for Text*. Switzerland: Springer International Publishing, 2018. ISBN9783319735306. [Online]. Available: https://doi.org/10.1007/978-3-319-73531-3_10

[71] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv Prepr. arXiv2203.05794*, 2022.

[72] L. McInnes, J. Healy, N. Saul, and L. Grosberger, "UMAP: Uniform Manifold Approximation and Projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, Sep 2018. [Online]. Available: https://doi.org/10.21105/joss.00861

[73] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," J. *Open Source Softw.*, vol. 2, no. 11, p. 205, Mar 2017. [Online]. Available: https://doi.org/10.21105/joss.00205

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org

[75] M. Grootendorst, "BERTopic," 2022. [Online]. Available: https://github.com/MaartenGr/BERTopic

[76] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008, vol. 39.

[77] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J.Wilson, S. Friedler, and S. Luccioni, "CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing," 2021. [Online]. Available: https://github.com/mlco2/codecarbon

[78] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling," in *Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, ser. SIGIR '13. Dublin, Ireland: Association for Computing Machinery, 2013. ISBN 9781450320344 pp. 889–892. [Online]. Available: https://doi.org/10.1145/2484028.2484166

[79] P. Suri and N. R. Roy, "Comparison between LDA & NMF for event-detection from large text stream data," in *2017 3rd Int. Conf. Comput. Intell. Commun. Technol.*, 2017, pp. 1–5. [Online]. Available: https://doi.org/10.1109/CIACT.2017.7977281