A News-Based Framework for Uncovering and Tracking City Area Profiles: Assessment in Covid-19 Setting

ALESSIO BECHINI, Department of Information Engineering, University of Pisa
ALESSANDRO BONDIELLI, Department of Computer Science, University of Pisa
JOSÉ LUIS CORCUERA BÁRCENA, PIETRO DUCANGE, FRANCESCO MARCELLONI, and ALESSANDRO RENDA, Department of Information Engineering, University of Pisa

In the last years, there has been an ever-increasing interest in profiling various aspects of city life, especially in the context of smart cities. This interest has become even more relevant recently when we have realized how dramatic events, such as the Covid-19 pandemic, can deeply affect the city life, producing drastic changes. Identifying and analyzing such changes, both at the city level and within single neighborhoods, may be a fundamental tool to better manage the current situation and provide sound strategies for future planning. Furthermore, such fine-grained and up-to-date characterization can represent a valuable asset for other tools and services, e.g., web mapping applications or real estate agency platforms. In this article, we propose a framework featuring a novel methodology to model and track changes in areas of the city by extracting information from online newspaper articles. The problem of uncovering clusters of news at specific times is tackled by means of the joint use of state-of-the-art language models to represent the articles, and of a density-based streaming clustering algorithm, properly shaped to deal with high-dimensional text embeddings. Furthermore, we propose a method to automatically label the obtained clusters in a semantically meaningful way, and we introduce a set of metrics aimed at tracking the temporal evolution of clusters. A case study focusing on the city of Rome during the Covid-19 pandemic is illustrated and discussed to evaluate the effectiveness of the proposed approach.

CCS Concepts: • **Computing methodologies** → **Information extraction**; • **Information systems** → *Data analytics*; *Web and social media search*; • **General and reference** → Metrics;

Additional Key Words and Phrases: City Areas profiling, online news clustering, NLP, text mining, streaming data, smart cities, Covid-19

This work was partially supported by the Italian Ministry of University and Research (MUR) in the framework of the CrossLab project (Departments of Excellence), and in the framework of PON 2014-2021 "Research and Innovation" resources - Innovation Action - DM MUR 1062/2021 - Titles of the Researches: "Progettazione e sperimentazione di algoritmi di federated learning per data stream mining" and "Modelli semantici multimodali per l'industria 4.0 e le digital humanities."
Authors' addresses: A. Bechini, J. L. C. Bárcena, P. Ducange, F. Marcelloni, and A. Renda, Department of Information Engineering, University of Pisa, Largo L. Lazzarino, Pisa 56122, Italy; emails: alessio.bechini@unipi.it, joseluis.corcuera@phd.unipi.it, {pietro.ducange, francesco.marcelloni, alessandro.renda}@unipi.it; A. Bondielli, Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3, Pisa 56127, Italy; email: alessandro.bondielli@unipi.it.

1556-4681/2022/07-ART125 $15.00
https://doi.org/10.1145/3532186

ACM Reference format:
Alessio Bechini, Alessandro Bondielli, José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, and Alessandro Renda. 2022. A News-Based Framework for Uncovering and Tracking City Area Profiles: Assessment in Covid-19 Setting. *ACM Trans. Knowl. Discov. Data.* 16, 6, Article 125 (July 2022), 29 pages.
https://doi.org/10.1145/3532186

1 INTRODUCTION

The Covid-19 pandemic has drastically affected people around the world, both in terms of how each one is facing his/her day-to-day life, and in terms of entire communities. For instance, the impact of restrictions adopted to mitigate the spread of the pandemic has recently been assessed in terms of urban crime [49], socioeconomic conditions [14], travel behavior [8], and waste management [55].

Different areas within the same city, some more than others, have experienced drastic shifts in how they have been lived, mainly due to rules concerning social distancing and prolonged lockdowns. For example, we can argue that areas around train stations have experimented an increase in criminality since a lower number of travelers pass along their streets. High-density and high-traffic areas may have instead experienced a decrease in road accidents and traffic jams because of a lower number of cars around. These shifts in urban life are an interesting aspect to monitor and profile for several reasons. First, the evaluation of the short-term impact of restrictions may provide a clearer picture of how the pandemic is actually affecting our lives. Second, the comprehension of long-term repercussions may be easier thanks to the development of appropriate tools to model changes in urban life. Third, the analysis of shifts may help local governments and security officers, who are in charge of managing public order and implementing forward-looking policies, to keep the city safe and livable.

Clearly, the problem becomes particularly interesting in the context of *smart cities*. The idea of a smart city actually hinges on the ability to exploit technological advancements, in terms of algorithms and available data, to provide improvements in the quality of life and available services for residents. The last few years have seen an ever growing interest in the development of systems and frameworks to monitor and analyze various aspects of life within cities. Technology is more than ever posed to play an important role in shaping the cities of the future. It can help local governments in taking informed decisions for the infrastructural development of the city, for handling problems related, for example, to crime and weather threats, and more in general for improving the quality of life of citizens. Research on smart cities has been steadily growing in the last few years, also because it encompasses several different domains of analysis and applications, such as the environmental, social, and mobility ones [23, 30, 62]. In this context, profiling city areas can be considered one of the main fields of application [31, 57].

Several approaches have been proposed in the literature for profiling of city areas [21]. One identified option relies on the use of structured data [22], setting up a framework to explore Web sources such as points-of-interest (e.g., restaurants, museums), traffic information, and house pricing. Concerning instead the use of more unstructured data such as local online newspaper articles, it has been proposed [15, 19] to exploit articles and tags to (i) identify a macro-categorization of news articles based on the semantic similarity between tags and (ii) classify news articles as belonging to one of such categories. The resulting information is used, on one hand, to cluster city areas based on the identified categories, and on the other hand to describe the various city areas in terms of the news reported for them. However, such approaches are not able to grasp the dynamic aspects of these phenomena, both in the case of generic topics like crimes and traffic, which are ordinarily covered by media, and in the case of new and emerging topics.

In this article, we aim to take into account the *evolution* of topics as well. Specifically, we want to propose a system for the automatic evaluation of modifications over time of the profiles of different city areas. To catch the main characteristics of the temporal evolution, we split continuous time into time windows and apply a purposely adapted streaming clustering algorithm for the identification of clusters of news articles in a given time window: this lets us uncover the differences between clustering outcomes obtained in adjacent windows. To track the evolution of clusters along subsequent windows, we introduce a set of metrics to describe the relation between a cluster in one window and clusters in the previous window based on the amount of shared members.

The evaluation of systems for profiling city areas is a rather complex task. The scope of the available research works is particularly broad, spanning different aspects and problems in the context of smart cities (See Section 2). Thus, to the best of our knowledge, no benchmark dataset is currently available to support direct evaluation and comparisons of models and analytical approaches. To overcome this limitation, we chose to quantitatively investigate the effectiveness of our framework via an experimental analysis on a paradigmatic case study. The experimental analysis has been carried out by considering news data for the city of Rome during 2020. The rationale behind this choice is twofold. First, it let us obtain a vast amount of news articles for a specific city, mostly geo-localized ones [15]; secondly, it represents an interesting evaluation ground for the proposed methodology. In fact, news related to the Coronavirus, which were obviously not present in previous years and in the very beginning of 2020, emerged in the subsequent months and were featured prominently, also obviously intertwining with other aspects of city life. This second aspect is particularly interesting, as it may allow us to build more robust and easy-to-model systems that can take into account unexpected variations in the data and enable a more in-depth understanding of how the pandemic affected life in the city. Concerning the choice of the specific city, it is worth underlining that our pipeline is highly versatile: Nowadays, there exist local online newspapers for any city, and this makes an analysis with fine spatial granularity plausible for any selected target area. We chose Rome as the area is familiar to most of the authors, thus making it easier spotting out possible problems or anomalous aspects during the system development.

The end goal of our work is to enable a *descriptive* and *fully unsupervised* analysis of changes over time in city area profiles. In our view, it is worth underlining two useful different aspects of the proposed framework. On one hand, the extracted knowledge can be exploited to automatically describe city areas in terms of what is reported to be happening in them in near-real time and across time. Moreover, the use of news articles and an unsupervised pipeline allows us to (i) avoid focusing on specific aspects of interest (e.g., crime rates, housing prices) and to (ii) discover novel descriptors for city areas as they appear over time, as in the case of the Coronavirus pandemic. On the other hand, the obtained results could proficiently support applications of different types. As a first example, we may consider a web mapping service, such as the popular Google Maps: Whenever users request a route planning, or explore a specific area, they can get aggregate information about what has recently featured that area; furthermore, this information is always up to date, and historical data series may also be available. Thanks to the information provided by our framework, users could decide to avoid a dangerous area characterized by numerous crimes and ask for alternative routes. As a second example, we can consider a real estate online platform or a physical agency: clearly, an updated characterization of city areas could represent an important strategic asset in this domain.

The main contributions of this article hinge on the design, development and deployment of a news-based framework, featuring the functionalities of news collection, data representation and processing, clustering, and knowledge extraction from clustering results. In particular, the novel data analysis pipeline entails the following contributions:

— The online news clustering task is addressed with a recently proposed density-based streaming clustering algorithm, adequately modified to automatically tune some of its parameters to adapt to evolving scenarios. Furthermore, an appropriate dimensionality reduction technique is employed to manage the high-dimensional real-valued vectors, i.e., *embeddings*, generated by language models to represent texts;
— A novel method for cluster labeling is proposed, aimed at revealing the topics covered in the news aggregated in the clusters, by leveraging a set of tags obtained from news articles;
— The identification of relevant patterns in the temporal evolution of clusters is addressed by introducing a novel set of metrics, defined to identify relationships among clusters in adjacent time windows;
— An in-depth experimental investigation is carried out with the proposed framework, leveraging a case study regarding the city of Rome during the Covid-19 pandemic, in order to evaluate the impact of the pandemic over the city in terms of clusters of reported news.

The rest of this article is organized as follows. In Section 2, the literature on city profiling and **Natural Language Processing** (**NLP**) is evaluated. Section 3 provides some background on specific techniques applied in the present work. Section 4 overviews the proposed system for city news clustering. Furthermore, Section 5 thoroughly describes the data preprocessing and clustering stage, whereas Section 6 describes the novel approaches for cluster labeling and tracking, which enable the knowledge extraction from clustering results. In Section 7, we evaluate our approach on a case study featuring news about the city of Rome during the Coronavirus pandemic, and we thoroughly discuss the obtained results. Finally, in Section 8, we draw proper conclusions and describe future directions.

2 RELATED WORKS

A significant amount of research has been carried out in subjects related to the system proposed in this article. The most relevant ones can be identified in the fields of smart cities frameworks and NLP techniques for text analysis.

2.1 Smart Cities: Overview and Frameworks

Research on smart cities and related applications has witnessed a constant growth in the last few years. Several definitions of "smart city" have been proposed in the literature [63]. Broadly speaking, a *smart city* is an urban area where disparate electronic methods are used to collect data, which, in turn, can be used to enhance services to the citizens (or create new ones) and improve their quality of life.

In this context, several different approaches have been proposed to exploit and analyze data collected from different sources in various urban areas, and to provide both administrations and citizens with strategic knowledge. Entire urban areas can be *profiled* on the basis of geo-referenced collected data. In this case, the goal of profiling refers to the ability to generalize on patterns and characteristics that are identified in the data. It is worth noticing that the body of data analytics works on profiling cities under various aspects is rather varied, with regard to the goals and scope of the research, and to the exploited data as well. In fact, data mining and machine learning techniques have been employed to characterize aspects of the city in very different domains, such as traffic and mobility [25], weather [42, 56], citizens' health [32, 51], economical factors [40], and crime prediction [1].

The vast majority of data-centric approaches proposed in the literature aimed either at investigating very specific aspects of city life, or at spotting out particular regions of interest. Fewer efforts have been put to provide end-to-end software systems for an automatic characterization of

city areas. Nevertheless, we can identify several interesting approaches. CityPulse [31] is a large-scale data analytic platform that exploits numerous web-based sources, such as social media posts and check-ins, to profile various aspects of city life. LocXplore [37] exploits various sources of geo-localized data, both unofficial (e.g., social media) and official (e.g., OpenData from authorities), to compare regions of interest within a specific city. The web application LiveCities [3] relies on location-based social media data and user profiling, with the purpose of visualizing recommendations of venues in a specific city.

The vast majority of approaches available in the literature do not directly factor the time dimension of the data in their analysis, i.e., how representations of cities (and city areas) change and evolve over time. This is in fact a challenging problem that requires to effectively model changes in distribution and characterization of the data (e.g., unpredictable trends in specific moments in time, emergence of new categories/classes).

A number of different sources have been exploited for various goals, such as mobile network activity [1], weather and temperature readings [51], energy consumption [4], and social media data. Other works exploit urban activities (e.g., Points-of-interests, citizen interactions) to feed recommendation systems [30, 38] and identify crucial urban activities [28, 50]. Social media data have shown to be particularly interesting, thus attracting a number of researches due to its informativeness in many different contexts. For example, authors in [34] use social media data to identify spatio-temporal patterns (e.g., mobility profiles). Location-based social media data have also been used to profile people behavior within cities [2], and to identify functional zones through the aggregation of tweets [18].

Most profiling systems, especially in the context of smart cities, usually take into account either structured or semi-structured data only. However, thanks to the advancements in NLP and *text mining*, textual data can be easily processed as well, and information of various nature can be extracted and modeled. In particular, newspaper articles represent an interesting target for NLP techniques, and moreover they have not been widely exploited in the context of smart cities. Newspapers have the advantage of reporting information on urban happenings almost in real-time. In this regard, the literature is quite scarce. Rivera et al. propose a classifier for RSS-like feed of news that is able to distinguish and locate traffic incidents, in order to timely alert citizens [53]. Similarly, Abid et al. propose to train a classifier to recognize events where some form of life loss happened (e.g., incidents, terrorism) from news [5]. Moreover, the authors of this article proposed a framework to classify news over a set of pre-determined categories (e.g., crimes, traffic), and used news to describe city areas [15]. The categories were obtained by grouping together similar tags from news with NLP and clustering techniques.

One drawback common to previously proposed approaches is represented by limitations in dealing with the modifications in collected data that may result in a different characterization of city areas. Even if plainly getting periodic snapshots may be sufficient in some settings, deeper insights into the evolution of urban areas may prove to be more useful, as well as the ability to recognize new and emerging elements not previously considered in the derived models. Our proposed approach is focused on improving the usefulness of city profiling frameworks with respect to this specific aspect.

We can easily realize that the known approaches to building city profiling frameworks differ in scope and end goals, types of employed algorithms, and data used. We can argue that such differences hamper any balanced comparison from the performance viewpoint, and the lack of publicly available source code for the proposed approaches worsen the scenario. Nevertheless, a comparison can be carried out from a functional perspective, and it has been reported in Table 1. The aspects that we deemed most relevant for the framework characterization include: data-related issues, such as the public availability of used data and the need for labeled data; algorithms used at

Table 1. Comparison of Available Frameworks

Framework	Scope/Goal	Data Sources	Data Availability	Data Labels Required	Employed Algorithms	Evolution/ Changes Tracking	Emerging Topic Discovery	Detection of Events
CityPulse [31]	City Profiling	IoT/Multimodal, Social Media	n/a	Yes	SensorSAX, CNN, CRF	Yes	No	Yes
LocXplore [37]	Region profiling	Social Media, Open Data	Open*	No	TextRank, ElasticSearch	n/a	No	No
LiveCities [3]	City Areas classification	Social Media	Open*	No	K-means	n/a	No	No
HotCity [38]	Recommender System	Social Media	Open*	No	n/a	Yes*	No	No
Bejar et al. [34]	Frequent Routes Discovery, User Behavor Profiling	Social Media	Open*	No	Frequent Itemset Mining, K-means	n/a	No	No
Zhong et al. [18]	City Areas profiling	Social Media	Open*	No	Topic Modeling, KNN clustering	n/a	Yes	n/a
Rivera et al. [53]	Traffic Incident Reporting	Online Newspapers	Open	Yes	SVM classifier	No	No	No
Abid et al. [5]	Incident Reporting	Online Newspapers	Open	Yes	Random Forest	No	No	No
Bondielli et al. [15]	City Areas profiling	Online Newspapers	Open	No	BERT	No	No	No
Proposed System	**City Areas profiling**	**Online Newspapers**	**Open**	**No**	**BERT, TSFDBSCAN.news**	**Yes**	**Yes**	**No**

Values marked with an asterisk are not clear-cut: For example, even if social media data can be generally considered as "open", their accessibility may be restricted by specific platforms and is subject to GDPR-like regulations.

training/inference time by the framework; the ability to track the evolution of monitored aspects across time, and understand how the evolution occurs; the ability to discover new topics as time progresses (e.g., emerging topics, new classes/categories); and the ability to perform detection of specific events.

2.2 Natural Language Processing and Text Mining Techniques

The development of neural language models allowed for the exploitation of distributional properties of texts to learn a numeric and semantic-aware representation of words and sequences [13, 24, 35, 45, 46, 52, 60], typically known as *word* or *sequence embeddings*. This last aspect, deemed crucial for many NLP tasks (e.g., sentiment analysis, machine translation) is particularly relevant in our setting, as we aim at grouping up articles with news on similar aspects of city life.

The vast majority of the approaches based on neural language models and vector representations of words and documents addresses problems by means of supervised learning schemes for solving downstream tasks. In particular, word embeddings and Language Models have shown to be able to obtain state-of-the-art performances in classification tasks. Conversely, only a handful of approaches attempted to extract knowledge from texts via unsupervised techniques such as clustering. Most traditional approaches in this regard exploit connectivity or centroid-based clustering approaches, such as Agglomerative Hierarchical clustering and K-Means clustering [15, 58]. Density-based approaches have not been fully explored yet, although they feature the desirable property of capturing arbitrarily shaped clusters with no prior assumption about the number of clusters. On the other hand, however, their performances consistently degrade as the number of features increases, and thus they are not necessarily suited to deal with the high-dimensional text representations obtained by language models. Several approaches have been proposed to alleviate this problem, either by exploiting algorithms more suitable to handle high-dimensional data, or by resorting to a preliminary dimensionality reduction. For example, a recent work [7] proposed to exploit clustering of document-level embeddings for topic modeling: the synergistic use of the UMAP [44] dimensionality reduction technique and the HDBSCAN [43] clustering algorithm led to significant results.

Similarly to many notable works in the field of document clustering [47], however, the authors in [7] considered a *static* setting, in which the whole collection of documents is available offline.

Nevertheless, when the goal is to extract knowledge from a continuously generated stream of news, the *evolving* nature of the dataset should be taken into account and suitable algorithms should be employed. The literature on clustering stream of documents is not as extensive as that for the static case. Evans et al. [27] conceived a prototype-based fuzzy clustering approach for processing web documents incrementally (i.e., in one pass) to endow search engines with the capability of grouping results based on their contextual similarity. Narang et al. [48] proposed to adopt an online clustering algorithm as a preprocessing step to find social discussion threads on the Twitter microblogging platform, by exploiting news media as an external source of information; however they focus on discovering discussion sequences leveraging social relationship information, and do not provide details about the clustering stage nor the algorithm adopted. Azzopardi and Staff [9] addressed the problem of grouping together news from disparate sources according to the event they describe; they use Bag-of-Words with **Term Frequency - Inverse Document Frequency** (**TF-IDF**) weighting scheme for text representation and a variation of the k-means algorithm to partition documents in an incremental manner. Recently, an approach for event-driven news stream clustering has been proposed [54]: authors combine dense and sparse numerical representation of documents based on BERT and TF-IDF scheme, respectively; furthermore, they show how the use of a proper fine-tuning objective and external knowledge (named "entity awareness") in pretrained transformer models help improve the clustering results. However, their focus is on the task of **Topic Detection and Tracking** (**TDT**) [6] and specifically on uncovering and segregating the *story chains* spawned by fine-grained real world events and echoed by correlated news articles, possibly originating from multiple redundant sources. Whereas in TDT context, a topic is closely associated with a real-world event, in turn, characterized by the triple (*location*, *time*, *people involved*), for our purposes a *topic* indicates the general subject discussed in the pieces of news. Furthermore, unlike [54], our focus is on city area profiling rather than TDT, and news from a single source are considered.

In the context of data stream mining, a fuzzy density-based clustering algorithm, named **Temporal Streaming Fuzzy DBSCAN** (**TSF-DBSCAN**) has recently been proposed [11]. Specific details on the algorithm are provided in Section 3.2. In this work, we adapt TSF-DBSCAN for the purpose of news stream clustering: the density-based approach ensures flexibility regarding the shape and the number of captured clusters; moreover, a fuzzy modeling fits the setting of document clustering, where the boundaries between clusters may not necessarily be sharp.

3 PRELIMINARIES

In this section, we first introduce some fundamental background on the topic of text representation, which is a crucial step of text mining pipelines. Then, we describe the TSF-DBSCAN algorithm, the reference clustering algorithm for our framework.

3.1 Text Representation Techniques: From Word Embeddings to Sentence-BERT

In the field of numerical representations of text, NLP research has made great leaps forward by identifying fixed-length word-level or sequence-level representations able to capture the semantics of words and sequences. We refer to word-level as representation that encode information for a single word, and to sequence-level as representation that encode information for a set of contiguous words, e.g., a sentence or a whole text. Such representations are often referred to in the literature as *distributed* representations of words and sequences [45], as they are typically learned by considering the *distributional* properties of texts, according to the distributional hypothesis [33], i.e., "words that occur in similar contexts tend to have similar meanings". Over the last years, a plethora of neural language models have been developed to this aim. Notably, the earliest neural approaches focused on building representations for single words, called *word embeddings* [13, 45, 46], learned

from corpora and stored for future use. The introduction of the Transformer Language Models [60] has then paved the way to *contextualized* representation of words. The main advantage of the transformer architecture and its implementations such as BERT [24] and GPT [17] with respect to other neural learning schemes (e.g., RNNs) is that they exploit an *attention mechanism* [60] to encode entire sequences of words at once, retaining information about the positioning of each word and modifying the output vector of the words based on its direct surroundings also at inference time. Thus, representations for words are not fixed. Moreover, Transformer language models allow for the pre-training and fine-tuning learning paradigm: first, the model is pre-trained on a general language learning task, and subsequently it is fine-tuned for downstream NLP tasks (e.g., classification).

For our purposes, the Sentence-BERT approach [52] is particularly interesting: It is a fine-tuning method that yields semantically relevant vector representations of sentences, which can be directly compared by means of measures such as *cosine similarity*. Notably, sequence level representations obtained by traditional pre-trained transformer models failed to model semantic correlation [24, 52]. To improve on these aspects, Sentence-BERT models are further tuned on sequence pair similarity and classification tasks based on the SNLI [16], MultiNLI [61], and STS benchmark datasets [20], aiming at maximizing the similarity measure of the resulting sequence embeddings for actually similar sentences. This, in turn, makes sequence-level embeddings suitable as input feature vectors for machine learning algorithms.

3.2 Density-Based Streaming Clustering: TSF-DBSCAN

TSF-DBSCAN [11] extends the popular DBSCAN algorithm; DBSCAN [26] identifies clusters by inspecting the neighborhood of each object x, on the basis of two parameters: the radius ε, which defines the neighborhood around x, and the minimum number *MinPts* of neighbors within the ε distance. Each object x with at least *MinPts* neighbors is marked as a *core object*. An object y that belongs to the ε-neighborhood of some core object x but does not satisfy the core condition is denoted as *border object*. Finally, an object that is neither a core nor a border object is marked as a noise object or an outlier. In a nutshell, TSF-DBSCAN adopts a two stage, online–offline, approach. During the online stage, the continuously collected objects are simply organized in adjacency lists. During the offline stage, the actual clustering procedure is executed as such. First, a forgetting mechanism is in charge of discarding outdated objects to comply with memory and computational constraint and to foster adaptation to non-stationary data distributions. In other words, objects collected before the current time window, with a user-defined size, are discarded. Then, a fuzzy adaptation of the DBSCAN algorithm is exploited to obtain a fine-grained partition of the most recent objects, not yet discarded by the forgetting mechanism. Fuzziness is introduced in the definition of the neighborhood of an object: In lieu of the crisp membership function adopted in the original DBSCAN and based on a distance threshold ε, TSF-DBSCAN exploits a fuzzy membership function based on two distance thresholds, ε_{min} and ε_{max}. Notably, only objects within ε_{min} are considered for the election of a core object. Nevertheless, an object may belong to the neighborhood of another object with a membership degree in the unit interval even if it is at a distance higher than ε_{min}, but lower than ε_{max} from the object itself. By relaxing the constraint on the neighborhood size, TSF-DBSCAN decouples two aspects of the DBSCAN clustering procedure: the identification of core objects and the membership assessment of border objects. As a result, it proved to be effective and efficient in modeling clusters with fuzzy overlapping borders, even in scenarios characterized by non-stationary data distributions. If the target objects are represented by news, and we want to *semantically* cluster them on the basis of their contents, a crisp approach is clearly unsuitable, as, in practice, an article may span multiple topics. Conversely, the use of fuzzy borders is likely more promising in modeling such a setting.

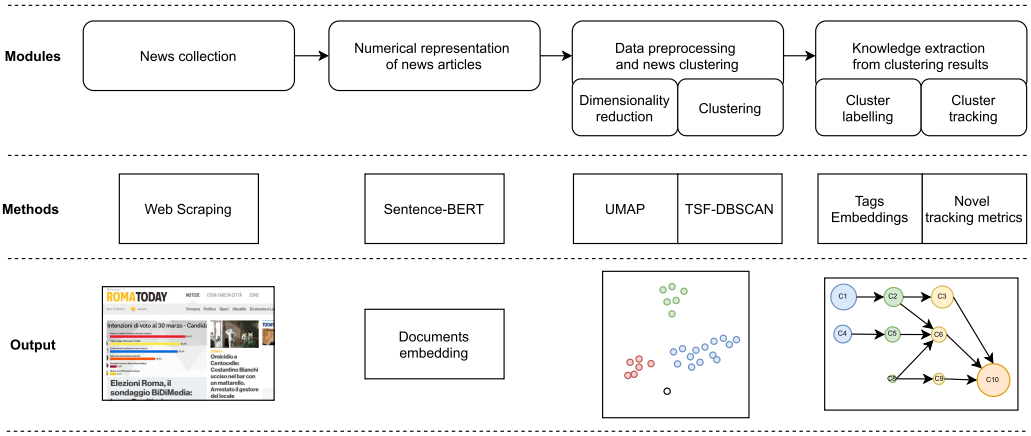


Fig. 1. High-level view of the proposed framework: software modules, tools, and relative output.

4 OVERVIEW OF THE PROPOSED FRAMEWORK

In this section, we provide a description of the proposed framework, focusing on the supported functionalities. It is worth recalling that we aim at exploiting data and metadata from online newspaper articles to profile city areas via an unsupervised approach. First, we collect newspaper articles from online sources. Then, we represent each article by leveraging a state-of-the-art Neural Language Model. Finally, we process and cluster articles for specific time frames in order to (i) identify groups of articles referring to the same general topic in a given time frame and (ii) track the evolution of the obtained clusters across time windows.

Figure 1 provides a high level view of the proposed framework. It encompasses four main modules: *news collection*, *distributed representation of news articles*, *news clustering*, and *knowledge extraction* from clustering results. For each module, the main methods used are indicated as well, along with the relative output. Each module is thoroughly described in the following.

News collection. The process of data collection from the web, and specifically from local news-oriented websites, can be easily applied for almost any possible target city, as this kind of information is readily available for the large majority of the cities worldwide. The end goal of this module is to obtain a coherent dataset out of the continuous stream of news articles in a given time span. Online news articles typically have an associated *timestamp*, i.e., the date of publication, and contain the *title*, a *summary* of the article, and the *body* of the article itself. For the sake of indexation, articles are often associated with a set of *tags* to describe the news contents. The news geo-localization can be given explicitly, or can be inferred. This kind of data can be easily collected either via the website API, if existing, or by programmatically scraping the news with robots.

Numerical representations of news articles. The capability of grouping up news articles on similar topics requires the use of text representations able to catch relevant semantic aspects. Specifically, news articles are entire text sequences that can be represented by single fixed-length real-valued vectors (a.k.a. "embeddings"). Despite the vast assortment of algorithms proposed to obtain meaningful embeddings, in the last years the neural language models emerged as the de-facto standard for word- and sequence-level embeddings and, among them, the Transformer architecture plays a prominent role.

In our system, we adopted a Sentence-BERT model [52], because of its ability to produce semantically meaningful feature representations of sequences, outperforming the standard BERT models [52].

We chose to encode only the title and the summary of the article, and not the body. We can argue that, due to how news are typically written, in most of the cases the key information regarding a news article can be found in the title and the first lines, or in this case the summary. Moreover, this let us also tackle two potential limitations of the Sentence-BERT models, namely the maximum input length and the method used to provide the sequence-level representation. The maximum input sequence length for the model we used is 512. Thus, articles longer than 512 words would be truncated nonetheless. In addition to this, the final sequence-level representation is obtained by averaging the word-level ones. Thus, the shorter the sequence, the more cohesive and representative its resulting vector is expected to be. Our choice, which is also backed by a set of preliminary experiments conducted on the different representations, may be regarded as the best tradeoff between encoded information and sequence length.

Data preprocessing and clustering. The numerical representation of news articles feeds a data mining module for data preprocessing and clustering. This module groups up news according to overarching topics, and needs to cope with some peculiar challenges of our setting: (i) the high-dimensionality of document embeddings (as obtained with Sentence-BERT), (ii) the evolving nature of the news stream, (iii) the presence of noise and uncertainty, e.g., an article may simultaneously belong to multiple clusters. As per the first aspect, a dimensionality reduction step is required, as all the components of the text embeddings are evenly relevant. Moreover, the number of clusters is not known in advance, and neither in the original nor in the reduced space any assumption can be made about the shape of the clusters. To address all these challenges, we developed *TSF-DBSCAN.news*, which is a streaming density-based clustering algorithm with incorporated dimensionality reduction and automatic parameter estimation. We used it to cluster groups of news articles for specific time frames with overlapping time windows. It extends the TSF-DBSCAN algorithm in our target context. A detailed description of the data preprocessing and clustering module is provided in Section 5.

Knowledge extraction from clustering results. The procedures in the previous module address the need of grouping up similar articles at any given time window, but further analysis is required for the comprehension of the clustering outcomes. The purpose of this module is twofold: cluster labeling, and cluster tracking. First, each cluster must be associated with an *interpretable label* that characterizes the most represented topics in the cluster. The cluster labeling sub-module exploits the tags of the articles to obtain such labels. Subsequently, we aim at monitoring and modeling the evolution of clusters along consecutive time windows. The cluster tracking functionality addresses this challenge and is conceived to assess the emergence, disappearance, evolution of the various topics over time. Section 6 is devoted to a detailed description of this module.

5 DATA PREPROCESSING AND CLUSTERING

TSF-DBSCAN addresses most of the requirements previously identified, but it still misses some crucial abilities. The scenario addressed in this article, in fact, is particularly hostile and precludes the plain application of an ordinary clustering algorithm for two main reasons: the high-dimensionality of the data, and the possible modification over time of the density of clusters. The *TSF-DBSCAN.news* has been developed to cope with these additional challenges.

Whenever objects are represented in a high-dimensional attribute space (as it is the case with word- and sequence-level embeddings), clustering algorithms struggle to get to significant results because of the so-called *curse of dimensionality* problem [39]. *TSF-DBSCAN.news* relies on the concept of density, which becomes less informative as the dataset dimensionality increases. In fact, the typical sparsity of data in high-dimensional attribute spaces makes all points appear similarly distant from one another. In other words, the distance of an object to the nearest object approaches the distance to the farthest one [12]. Consequently, the concept of local neighborhood, which determines what objects are *core* ones, turns out to be inappropriate to drive the cluster analysis. An adequate solution to this problem can rely on some form of dimensionality reduction as a preprocessing step, to make the downstream clustering procedure operate in a lower dimensional space. It has been shown [41] that a manifold learning technique named **t-distributed Stochastic Neighbor Embedding** (**t-SNE**) [59] can be used for this purpose since, in general, the number of disjoint clusters in the target space coincides with the number of disjoint clusters in the original space. In our work, we consequently revisit the TSF-DBSCAN algorithm: During the online stage, the embedding vector of each collected article is computed; when the offline stage is triggered (i.e., a periodic condition is met) we use **UMAP** (**Uniform Manifold Approximation and Projection** for Dimension Reduction) [44], a recent enhanced version of t-SNE, to project the article level embeddings in a lower dimensional space. The main advantage of UMAP over t-SNE is that it is able to keep the global structure of the data, and to better scale up on large datasets [36].

As for dealing with evolving scenarios, the original TSF-DBSCAN is inherently capable to handle non-stationary distributions, such as those characterized by the emergence, disappearance, or gradual movement of clusters. However, it is not flexible enough to adapt to changes over time of the density of clusters, since it exploits an initial *static* configuration of the parameters ε_{min} and ε_{max}. Such a variation in cluster density can derive from *concept drift* [29, 64], i.e., the non-stationarity of the data generation process, and it can likely afflict the stream of news, e.g., as a result of a burst of news on emerging "hot" topics. Furthermore, it is widely acknowledged that also in the original DBSCAN a very critical issue is the choice of the ε parameter [26], as it defines the neighborhood extent for any object, therefore implicitly affecting the number and shape of the detected clusters: Thus, the ability to update this parameter over time can provide a certain level of flexibility to the framework. For accommodating the evolving nature of news streams, in the proposed *TSF-DBSCAN.news,* we included a stage for the automatic tuning of the threshold parameters prior to each reclustering step, based on a recently published approach [10]: the statistical modeling of the density distribution of objects lets us shape a heuristic to estimate ε_{min} and ε_{max}. The underlying idea recalls the original proposal of the authors of DBSCAN [26] to resort to the *k-dist* function, which associates each object in the dataset with the distance from its kth nearest neighbor. However, instead of requiring the user to manually select the distance threshold upon the visual analysis of the resulting *k-dist* plot, the heuristic automatically derives the parameter values for the fuzzy membership function on the basis of a Gaussian Mixture modeling of the *k-dist* array. Notably, the effectiveness of the approach has been previously shown on a number of synthetic benchmark datasets [10], also in presence of artificially added noise.

Figure 2 shows the adapted version of the clustering algorithm exploited in this article, highlighting the steps of dimensionality reduction and adaptive tuning of threshold parameters.

The outcome of this module for the generic time window t is the set P^t of the identified disjoint clusters, informally named "partition", along with the membership of each object to a cluster in P^t (or its recognition as an outlier). Notably, a defuzzification process may be applied on the output of TSF-DBSCAN so that each object is assigned to at most one cluster based on the highest membership degree.

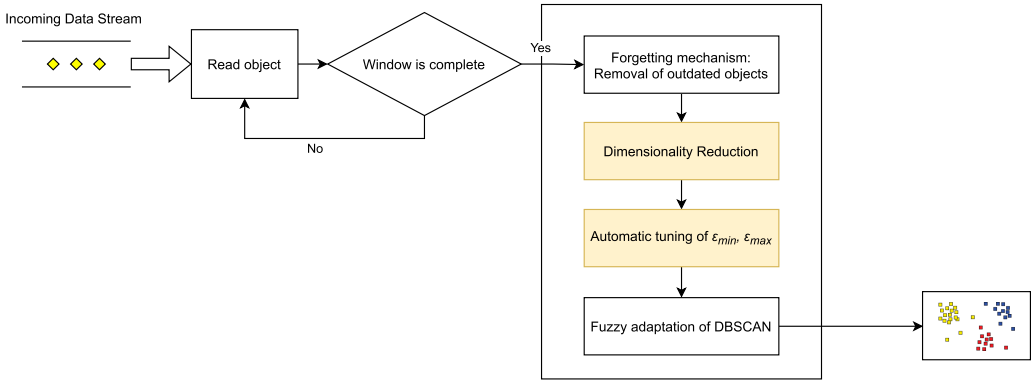


Fig. 2. Main steps of the data preprocessing and clustering module: yellow blocks indicate the novel components introduced in *TSF-DBSCAN.news* w.r.t. the original TSF-DBSCAN.

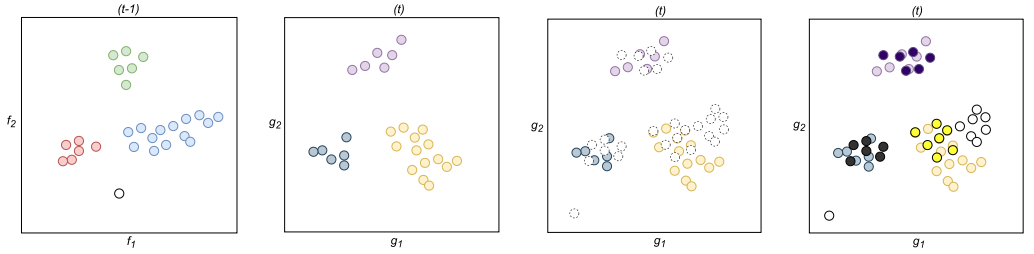
6 KNOWLEDGE EXTRACTION FROM CLUSTERING RESULTS

The last module of the proposed framework is in charge of extracting knowledge from the outcome of the previous clustering step. The activities to be carried out are the cluster labeling and cluster tracking, supported by dedicated sub-modules.

6.1 Cluster Labeling

The procedure proposed for cluster labeling is based on topic modeling, by exploiting embeddings for a semantics-aware comparison of possible labels [7]. The labeling of any cluster can be carried out on the basis of the tags associated with the composing articles. As tags are standalone words, no context-sensitive representation is required for them (Sentence-BERT is unreasonably too complex for this task). Thus, a standard word embeddings model has been chosen, and specifically *fastText* [13, 35], which encodes each tag in a 300-dimensional space. An accurate yet flexible labeling can be based on a *reference tag dictionary*: only the tags in this collection can be used for labels. Such a dictionary must contain a curated set of well-defined relevant tags, along with tags extracted from the most recent articles, which are possibly able to describe emerging topics. In the approach adopted for the definition and maintenance of the reference tag dictionary, we initially collect the article tags for a given period of time; this set is first filtered according to a minimum frequency threshold for the occurrences of each tag, and subsequently the tags that are not deemed sufficiently informative are removed and placed in a dedicated blacklist. For example, in our case study, "Rome" and other tags concerning locations of the city are not sufficiently informative for our purposes. As new articles arrive, the frequent relative tags not already present in the blacklist are temporarily added to the dictionary: periodically, a revision of the dictionary is performed to decide whether to definitely include the new tags in the set of relevant ones.

Let t denote a generic time window. The labels to associate with each cluster C_i^t are generated as follows. For each time window t, the clustering algorithm generates a partition $P^t = set(C_i^t)$; First, for each cluster C_i^t in P^t, we identify the top-K ($K = 10$ in our experiments) most frequent tags along with their frequencies; if they are not present in the *reference tag dictionary*, the *N/A* label is assigned. Second, we obtain the distributed representation of each frequent tag (i.e., its *tag embedding*) according to the pre-trained fastText model. Third, we compute the *centroid embeddings* $\overrightarrow{WTag_i^t}$ as the component-wise weighted average of the tags embeddings in each



(a) P^{t-1}: Output partition at window $(t-1)$

(b) P^{t}: Output partition at window (t)

(c) Objects of window $(t-1)$ are projected according to the transformation adopted for window (t) and represented with a dashed line.

(d) *Virtual assignment*: objects of window $(t-1)$ are virtually assigned to the clusters discovered at window (t) and represented with brighter colors.

Fig. 3. Schematic representation of the preliminary steps for the cluster tracking procedure for a simplified bi-dimensional case. (f_1, f_2) and (g_1, g_2) indicate the dimensions on which objects are projected in time window $(t-1)$ and (t), respectively. Objects not assigned to any cluster are reported in white. Best viewed in color.

cluster C_i^t; tag embeddings are weighted by their frequency:

$$\overrightarrow{WTag}_i^t = \frac{\sum_{k=1}^{K} f_{i,k} \cdot \overrightarrow{tag}_{i,k}}{\sum_{k=1}^{K} f_{i,k}}, \tag{1}$$

where $\overrightarrow{tag}_{i,k}$ is the embedding representation of the kth most frequent tag in cluster C_i^t, $f_{i,k}$ represents the number of occurrences of $\overrightarrow{tag}_{i,k}$, and the summation notation indicates component-wise summation.

Finally, we identify the nearest neighbor of $\overrightarrow{WTag}_i^t$ in the *reference tag dictionary* with respect to cosine similarity. Such a tag gives an indication on the principal topic for the cluster. Optionally, to grasp the nuances of cluster traits, a label may consist in an ordered sequence of tags: In our case study, we will report the three nearest neighbors.

6.2 Cluster Tracking

The clustering algorithm computes an output partition from scratch at each time window. Therefore, investigations on the evolution of profiles ask for a method to track clusters across consecutive time windows. The proposed method hinges on the definition of three metrics that characterize the *purity*, *coverage,* and *preservation* of clusters. The joint assessment of these metrics allows us to capture salient phenomena concerning the partition evolution, including emergence and merging of clusters, or whether and up to what extent a cluster is maintained over time.

In order to evaluate the proposed metrics, we must first devise a procedure to identify the relationship between objects/clusters of two consecutive windows. It is schematically depicted in Figure 3: After clustering generation at time t, objects of the previous window $(t-1)$ are projected into the current reduced space determined according to the dimensionality reduction transformation learned on the data in time window t. Then, the membership to clusters in P^t of each projected object is evaluated, applying the criteria used in the adopted clustering algorithm for determining whether an object belongs to a cluster: In the following, this procedure is referred to as *virtual assignment*.

In the following, we first introduce the notation and then we define the *metrics to support cluster tracking* across subsequent time windows.

Let C_i^{t-1} and C_j^t be two generic clusters in partitions P^{t-1} and P^t obtained for time windows $t-1$ and t, respectively. Using the notation $|\cdot|$ for set cardinality, the number of clusters in a generic window t corresponds to $|P^t|$. Furthermore, we denote with $C_{i\to j}$ the set of objects in C_i^{t-1} that are virtually assigned to cluster C_j^t according to the procedure illustrated in Figure 3.

***Purity* of C_j^t.** This measure of disorder can be applied in case at least one object of a cluster in P^{t-1} is virtually assigned to C_j^t. If this condition holds, we can indicate with $Q_j^{t-1} = \{k \in P^{t-1} \mid C_{k\to j} \neq \emptyset\}$ the subset of the clusters in partition P^{t-1} with at least one object virtually assigned to C_j^t. Members of Q_j^{t-1} are referred to as "contributing clusters". The index is computed by exploiting the concept of *normalized entropy* for a cluster C_j^t, defined as

$$E_j^t = -\sum_{i\in Q_j^{t-1}} \frac{p_{i\to j}\cdot \log(p_{i\to j})}{\log(|Q_j^{t-1}|)}, \tag{2}$$

where $p_{i\to j}$ is the fraction of objects in $C_{i\to j}$ w.r.t. all the objects from Q_j^{t-1} virtually assigned to C_j^t, i.e.,

$$p_{i\to j} = \frac{|C_{i\to j}|}{\sum_{k\in Q_j^{t-1}} |C_{k\to j}|}. \tag{3}$$

The normalized entropy ranges between 0 and 1. Thus, we can define the purity of cluster C_j^t as

$$Pur_j^t = 1 - E_j^t. \tag{4}$$

A high value of purity for cluster C_j^t occurs whenever the objects from the previous window that are virtually assigned to C_j^t largely originate from one single cluster. Conversely, a provenance evenly involving multiple clusters yields a low purity value.

***Coverage* between C_i^{t-1} and C_j^t.** Coverage is defined as follows:

$$Cov_{i\to j}^t = \frac{|C_{i\to j}|}{\max(|C_j^t|, |C_{i\to j}|)}. \tag{5}$$

The coverage ranges from 0 to 1, with 1 testifying that the number of objects belonging to $C_{i\to j}$ is equal to, or higher than, the number of objects in C_j^t. Incidentally, the cluster l in time window $t-1$ that most contributes to C_j^t can be found as $l = \underset{i\in P^{t-1}}{\text{argmax}}(Cov_{i\to j}^t)$.

***Preservation* of C_i^{t-1} in C_j^t.** This metric evaluates the fraction of objects in cluster C_i^{t-1} that are virtually assigned to cluster C_j^t. Preservation is defined as

$$Pre_{i\to j}^t = \frac{|C_{i\to j}|}{|C_i^{t-1}|}. \tag{6}$$

The preservation ranges between 0 and 1. Preservation has its maximum value when all the objects in cluster C_i^{t-1} are virtually assigned to the same cluster C_j^t.

Given two consecutive partitions P^{t-1} and P^t, some notable patterns, which we name *Continuity*, *Topic Emergence*, *Topics Fusion*, and *Topic Expansion*, can be described in terms of the metrics defined above. Formally:

Continuity: Let $l = \underset{i \in P^{t-1}}{\text{argmax}}(Cov^t_{i \to j})$.
WHEN Pur^t_j is *high*
AND $Cov^t_{l \to j}$ is *high*
THEN C^t_j originates from C^{t-1}_l.
Topic Emergence: Let $l = \underset{i \in P^{t-1}}{\text{argmax}}(Cov^t_{i \to j})$.
WHEN $Cov^t_{l \to j}$ is *low*
THEN C^t_j represents an *emerging topic*.
Topics Fusion:
WHEN $\sum_{i \in P^{t-1}}(Cov^t_{i \to j})$ is high
AND $\sum_{i \in P^{t-1}}(Pre^t_{i \to j})$ is high
THEN C^t_j is the *fusion of different clusters* from P^{t-1}.
Topic Expansion:
WHEN $Pre^t_{i \to j}$ is *high*
AND $Cov^t_{i \to j}$ is *low*
THEN C^t_j is expanding from C^{t-1}_i.

Although other different patterns could be envisaged, we identified these ones as the most significant for our news stream investigations: In the following section, we show how they occur and are used in our monitoring campaign. Notably, uncovering the occurrence of one of these patterns is an important step in cluster evolution analysis, but it should be always complemented by information from the labeling of the involved clusters. Furthermore, it is worth underlining that the adjectives *high* and *low* provide a rough indication: *automatic* pattern matching requires to specify appropriate threshold values.

7 SYSTEM DEPLOYMENT AND CLUSTERING RESULTS: A CASE STUDY ON THE CITY OF ROME

The proposed framework exploits a data processing pipeline that has been developed in Python. To evaluate and demonstrate it usefulness, we set up a case study regarding the city of Rome. In the following, we first describe the experimental setup, with details on the dataset and on the system configuration; subsequently, we report and discuss the results of our experimental campaign.

7.1 Dataset Description

The reference dataset for our investigation is extracted from RomaToday,[1] a well-known online newspaper with news regarding exclusively the Italian Capital and its direct surroundings. An example of the relative article data and metadata is reported in Table 2. We collected news in a monitoring campaign from June 2019 to June 2020, roughly centered over the initial spread of Covid-19 in Italy, with the intent to catch its impact on the profiles of the city areas. The dataset contains about 15,000 news articles; Table 3 summarizes its main statistics. Notably, a non-negligible number of articles have no associated tag at all (1,202 out of 15,214), and this indicates how a categorization based on other elements of the article, such as the title and the summary, might be important in practice. Figure 4 shows the top-25 most frequent tags in the whole monitored period, providing an indicative picture of the most characterizing topics covered in our dataset. Moreover, the spatial distribution of the 2020 articles is reported as a heatmap in Figure 5; a grid

[1] www.romatoday.it.

Table 2. Raw Data Extracted from News Articles

Type	Sample value
URL	https://www.romatoday.it/attualita/coronavirus-test-sierologici-polizia-locale.html
Time Stamp	2020-04-25 09:27:24
Title	Coronavirus, serological tests also to the Local Police: the Lazio Region pays
List of Tags	Coronavirus
Location	Garbatella, Via Rosa Raimondi Garibaldi
Summary	The Region replies to the protests of the May Commander who had made himself available, in the absence of contributions, to bear the costs for the serological tests to the Local Police officers.
Body	The Lazio Region will finance the costs of performing serological tests also for members of the Local Police of Rome Capital. Like health workers and men who report to other law enforcement agencies, the police will also be able to benefit from it...

Original texts are in Italian; only the English translation is reported here.

Table 3. Summary of the Reference Dataset: News Extracted in the Whole Monitoring Period

Data Attribute	Value
Number of articles	15,214
Number of non-geolocated articles	7,614
Number of articles with no tag	1,202
Number of different tags	7,962
Avg. number of tags per article	2.341
Min. number of tags per article	0
Max. number of tags per article	14
Avg. number of articles per month	1,144
Min. number of articles per month	950
Max. number of articles per month	1,288

with 1-km step is used over the urban area. Unsurprisingly, the highest density of news is present in downtown areas.

7.2 Configuration of the Framework Parameters

The numerical representation of news articles represents one of the key components of the proposed system. As per the choice of the language model, a variety of pre-trained Sentence-BERT models is available as part of the `sentence_transformers` library.[2] Obviously, the choice should depend on the language used in the target news articles. Our case study concerns Italian newspapers, but, unfortunately, no Sentence-BERT model specifically tuned on Italian data is currently available; thus, we chose to consider a multilingual model (i.e., `xlm-r-bert-base-nli-stsb-mean-tokens`) as it can provide out-of-the-box reliable representations of sequences in a variety of languages.

We encode the title and the article summary with the Sentence-BERT model, and obtain the 768-dimensional article-level embeddings. Conversely, for tag embeddings, the fastText model for Italian has been used in the case study.

The parameters of the algorithms used in the proposed framework have been set according to the indications of the specialized literature. UMAP has several hyper-parameters that affect the dimensionality reduction operation, and the most important ones are the number of neighbors,

[2] https://github.com/UKPLab/sentence-transformers.

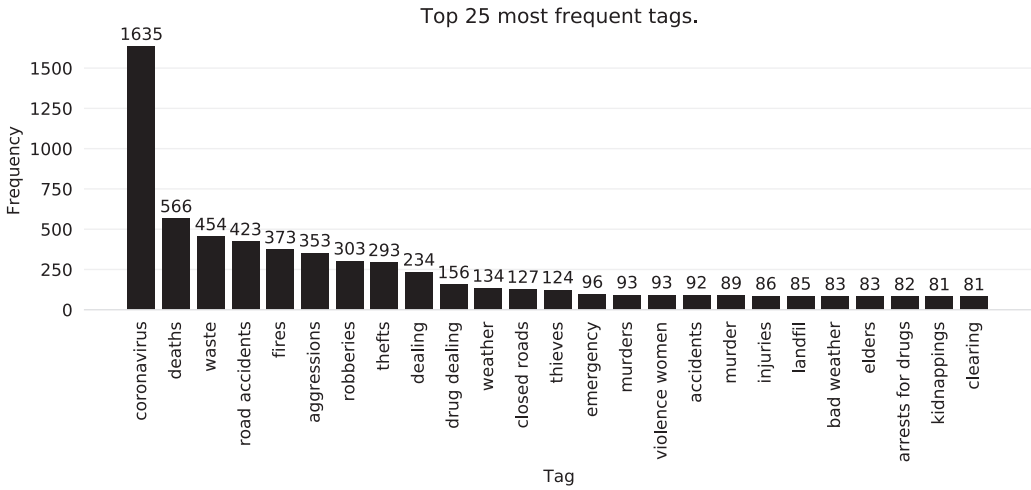


Fig. 4. Top-25 most frequent tags extracted from the dataset.

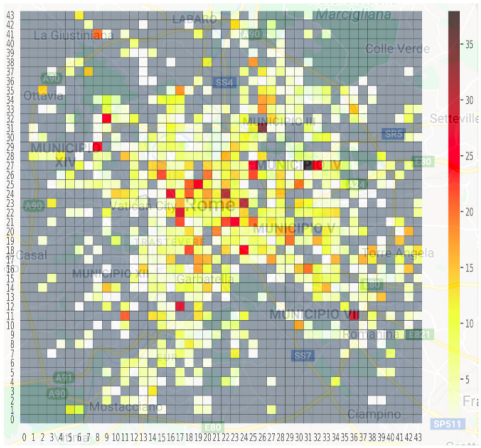


Fig. 5. Spatial distribution over the Rome metropolitan area of the news in the reference dataset.

the used metric, and the dimensionality of the reduced space. The number of neighbors controls the balance between preserving global structure and local structure in the reduced data space: the larger the value of this parameter, the larger the emphasis on global structure preservation [44]. We executed UMAP with the recommended parameter setting (30 neighbors) to avoid noisy fine grained clusters. As for the metric, we chose cosine distance as it is considered a standard metric and widely exploited in NLP tasks to effectively measure the distance between word- and document-level embeddings. Furthermore, we projected the 768D Sentence-BERT embeddings in a lower dimensional feature space to make the downstream cluster analysis more effective. We set the dimensionality of the target space as 5, following the indications of the related literature [7]. The *TSF-DBSCAN.news* algorithm was executed with $MinPts = 5$, while the values of the distance thresholds ε_{min} and ε_{max} were automatically derived at each reclustering step, as described in Section 5. The offline stage of *TSF-DBSCAN.news* was evaluated with a period of one month. Furthermore, we tuned the forgetting mechanism so that only the news articles collected in the

Table 4. Summarized Clustering Results of Period June 2019 to June 2020

Time window	#objects	#clusters	#objects per cluster		#outliers
			Max	Min	
June 2019	1,567	6	1,478	6	0
July 2019	1,575	4	1,134	10	7
August 2019	1,220	11	503	10	0
September 2019	1,340	9	812	7	2
October 2019	1,548	16	765	7	11
November 2019	1,411	4	915	15	0
December 2019	1,342	4	916	6	0
January 2019	1,247	13	521	6	0
February 2020	1,377	10	787	5	2
March 2020	1,427	3	1,397	6	0
April 2020	1,313	23	286	8	12
May 2020	1,424	25	378	4	25
June 2020	1,552	24	317	7	9

previous five weeks, approximately, were considered as input for each reclustering step. In other words, we allowed an overlap of one week between two consecutive evaluations of the offline step of the clustering algorithm.

7.3 General Results

Table 4 shows some summarized statistics of the clustering results over the various time windows. Specifically, for each window, we report the number of articles, the number of clusters, the maximum and minimum number of articles per cluster, and the number of outliers. We can observe that the number of clusters generated at each time window ranges from a minimum of just three clusters in March 2020 to more than 20 clusters in the final period of our monitoring campaign. Such a flexibility of the density based clustering algorithm with respect to the number of discovered clusters let us model some interesting patterns; in fact, the topic of Covid-19 pandemic almost monopolized the online news in March 2020, while a substantial fragmentation of topics characterized the subsequent months.

Notably, the clustering operation spotted out also some outliers in several time windows. For instance, in July 2019 four clusters were identified in total, with computed labels referring to the topics "waste-emergency", "deaths", "heat", and "thefts". The outliers showed no similarity with such clusters, and their tags referred to "cubs", "scouts", and "people exploitation".

Providing an overall effective representation of the clustering results is not trivial. Figure 6 shows a stacked bar plot that represents the main clusters discovered in the central period of our analysis, between December 2019 and May 2020: the height of each bar is proportional to the number of objects in the cluster. Furthermore, each cluster bar is labeled with the three nearest neighbors to $\overrightarrow{WTag}_i^t$, as defined in Equation (1), along with their cosine similarity w.r.t. $\overrightarrow{WTag}_i^t$. For visualization purpose only, clusters with less than 100 objects are grouped and labeled as *others*. A visual analysis of the figure indicates the presence of some dominant and recurring topics, with several clusters related to the tags "waste emergency" or "thefts" and "robberies". Distinctly, this trend has been drastically disrupted by the emergence of the "coronavirus" topic. However, to get a more thorough understanding of the situation and to model the evolution of clusters over time, we resort to the newly defined cluster tracking metrics.

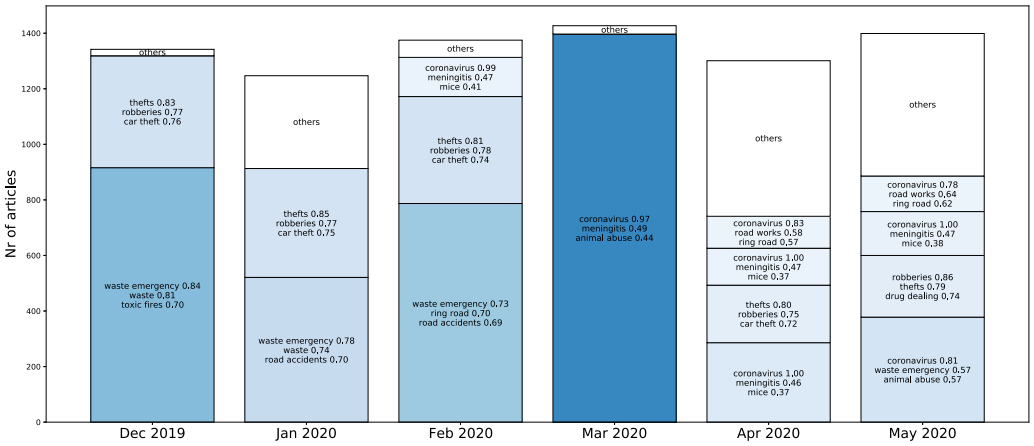


Fig. 6. Clustering results in the period December 2019–May 2020. Each cluster is labeled using the procedure defined in Section 6.1: each tag is shown along with the cosine similarity between its embedding and the centroid embedding $\overrightarrow{WTag}_i^t$. Color saturation is proportional to the number of objects in a cluster. For the sake of clarity, clusters with less than 100 objects are collectively indicated as *others*.

7.4 Metrics and Pattern Evaluation

In this section, we provide some examples of evaluation of the metrics defined in Section 6.2. Table 5 reports the computed values of these metrics for the cluster identified by the tags "waste emergency", "waste", and "road accidents" from the January 2020 time window. The high value of *Purity* ($Pur = 0.745$) suggests that most of the objects from the previous time window that are virtually assigned to this cluster come from one single cluster: In our case, 597 out of the 650 objects assigned to the current cluster (more than 90%) originate from the same cluster. The source cluster, identified by the tags "waste emergency", "waste", and "toxic fires", is fairly preserved in the current cluster ($Pre = 0.652$) since 597 out of 916 objects are mapped therein. In addition, the number of objects projected from such a source cluster (597) is higher than the total number of objects of the current cluster (521). Thus, according to Equation (5), we can report a *Coverage* of 1.

Such high values for *Coverage* and *Purity* indicate a match for the *Continuity* pattern: the current cluster *originates* from a cluster from the previous window. Indeed, they share the tags "waste emergency" and "waste" with high cosine similarity values. Although other clusters from the previous time window also have objects mapped to the cluster under investigation, their *Coverage* values are very low, thus indicating no significant pattern match.

Another example of the computed metrics is provided in Table 6, related to the cluster labeled by "thefts", "robberies", and "car theft". Also in this case, an occurrence of the *Continuity* pattern is identified over a cluster from the previous time window that shares similar topics, and features exactly the same tags. Although it is evident that most of the examples projected onto the current cluster originate from one single cluster of the previous time window (348 out of 404), the *Purity* value ($Pur = 0.419$) is significantly lower than in the example discussed before. In the following, we empirically show that this value should nevertheless be considered as *high*. Since the objects projected onto the current cluster originate from *two* clusters only, we computed the purity value of all the possible pairs of integers that sum up to 404. Figure 7 reports the CDF of such array of purity values: The shape of the plot denotes a strong skewness of the distribution of purity values, and therefore the value found for the analyzed cluster, very close to the 75th percentile (0.456), can be considered relatively high.

Table 5. Cluster Tracking Metrics for the "Waste Emergency"-Related Cluster in Time Window January 2020

Waste-related cluster - Jan. 2020				
Label	$\lvert C_j^t \rvert$		**Purity**	
waste emergency: 0.777 waste: 0.741 road accidents: 0.695	521		0.745	
Relative contributing clusters				
Label	$\lvert C_i^{t-1} \rvert$	$\lvert C_{i \to j} \rvert$	**Cov.**	**Pre.**
waste emergency: 0.839 waste: 0.809 toxic fires: 0.702	916	597	1	0.652
thefts: 0.827 robberies: 0.773 car thefts: 0.760	403	34	0.065	0.084
weather: 0.874 weather threats: 0.821 bad weather: 0.642	17	13	0.025	0.765
trees: 0.713 boars: 0.661 fallen trees: 0.651	6	6	0.012	1
Sum:		650	1.102	2.501

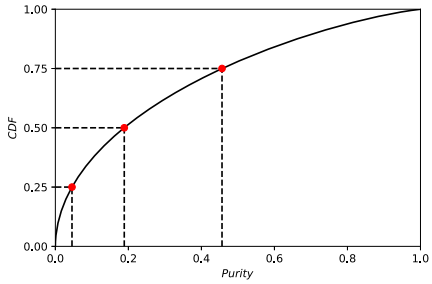


Fig. 7. Empirical **cumulative distribution function** (**CDF**) of purity for all the possible pairs of integers with a given fixed sum value.

Finally, Table 7 reports the metric results for the cluster identified by the tags "coronavirus", "deaths", and "meningitis". In this case, we can observe that very few objects from the earlier window are projected there, leading to a low maximum value of *Coverage*. Thus, an occurrence of the *Emerging topic* is present: it corresponds, in early 2020, to the first news related to the Coronavirus pandemic.

7.5 Impact of Covid-19 on News Clusters

In order to evaluate the impact of the Coronavirus pandemic and how it reflects on our data and on reports for the city of Rome, we recall here several important dates related to the outbreak and evolution of the pandemic. According to the **World Health Organization** (**WHO**), on December 31st, 2019, several health authorities around the world contacted WHO seeking additional information about this "viral pneumonia". However, it was not until January 31st, 2020, that the

Table 6. Cluster Tracking Metrics for the "Thefts"-Related Cluster in Time Window January 2020

Theft-related cluster - Jan. 2020		
Label	$\lvert C_j^t \rvert$	**Purity**
thefts: 0.846 robberies: 0.765 car theft: 0.755	521	0.419

Relative contributing clusters				
Label	$\lvert C_i^{t-1} \rvert$	$\lvert C_{i\to j} \rvert$	**Cov.**	**Pre.**
thefts: 0.827 robberies: 0.773 car theft: 0.76	403	348	0.887	0.863
waste emergency: 0.839 waste: 0.809 toxic fires: 0.702	916	56	0.143	0.061
Sum:		404	1.03	0.925

Table 7. Cluster Tracking Metrics for the "Coronavirus"-Related Cluster in Time Window January 2020

Coronavirus-related cluster - Jan. 2020		
Label	$\lvert C_j^t \rvert$	**Purity**
coronavirus: 0.951 deaths: 0.530 meningitis: 0.498	28	0.000

Relative contributing clusters				
Label	$\lvert C_i^{t-1} \rvert$	$\lvert C_{i\to j} \rvert$	**Cov.**	**Pre.**
waste emergency: 0.839 waste: 0.809 toxic fires: 0.702	916	4	0.143	0.004
thefts: 0.827 robbery: 0.773 car theft: 0.760	403	4	0.143	0.009
Sum:		8	0.285	0.014

first two positive cases of Coronavirus in Italy were reported at the Spallanzani hospital in Rome. As time elapsed, more cases were reported, and the first localized lockdown was declared in the Lombardia region on February 23rd. On March 9th, a country-wide lockdown was put into action, and kept until May 18th. In the bottom part of Figure 8, it is depicted the popularity of the search query "*coronavirus*" in Google Search,[3] and the mentioned dates have been clearly marked with red lines. It is quite evident that the peaks in the chart are associated with the events at the marked dates and, in particular, with the worsening of the pandemic situation. In the following, we take an in-depth look at how our system can help understand the evolution of clusters over the period of the pandemic diffusion.

We focus on the three most relevant topics, namely "Waste", "Theft", and "Coronavirus", and identify the clusters whose labels belong to these broad semantic areas. Figure 8 (top) shows how

[3]https://trends.google.com/trends/explore?date=2019-10-25%202020-05-31&geo=IT-62&q=coronavirus.

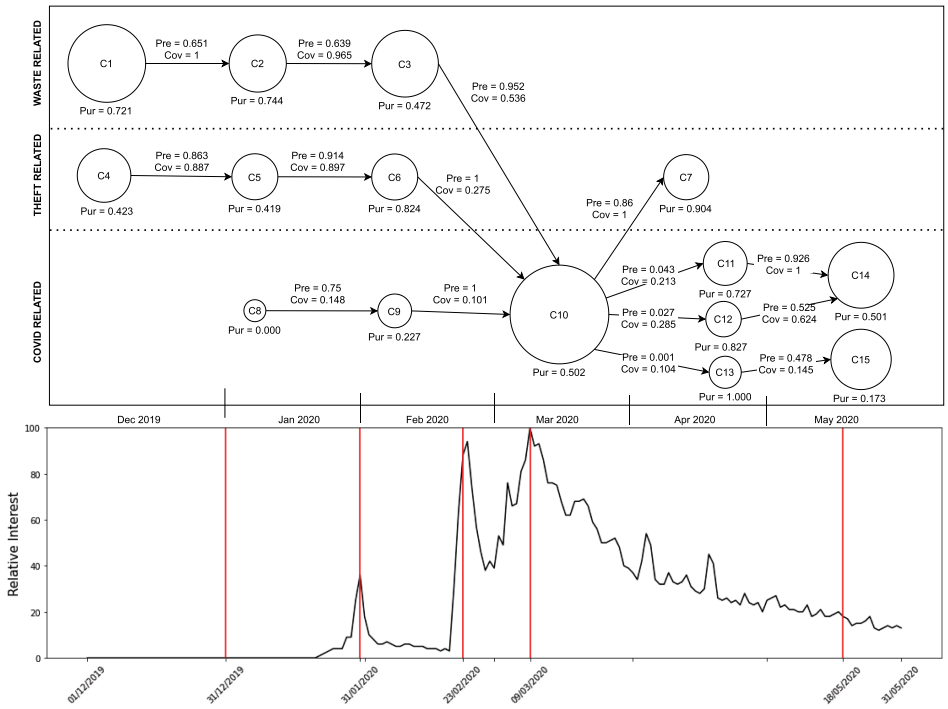


Fig. 8. (Top) Clustering evolution and frequency of news for the main macro-topics from December 2019 to May 2020. Each cluster, identified by an ID, is represented along with its *Purity* value; the relative tags are listed in Table 8. The size of each circle is proportional to the cluster cardinality; *Preservation* and *Coverage* values are reported for the most relevant links. (Bottom) Google search trend for the keyword "*coronavirus*" during the same period. Some relevant dates (discussed in Section 7.5) are indicated by a red vertical line.

such topics evolve from December 2019 to May 2020 by representing the most relevant clusters and the links between them along with the values of the metrics introduced in this article. Table 8 reports the labeling of the clusters involved in the analysis.

The first cluster labeled as "Coronavirus"-related (i.e., C8) coincides with the period with the first-two positive cases reported in Italy. In the previous section we have shown, by analyzing Table 7, that this cluster represents the emergence of the coronavirus topic. Clearly, as the pandemic evolved, also our identified clusters evolved accordingly. For example, clusters C9 and C10 roughly coincide at least temporally with the Lombardy and Nation-wide lockdowns, respectively. Furthermore, if we look at the values for *Preservation* and *Coverage* for the clusters, they match the *Topic Expansion* pattern. Moreover, clusters related to "Waste" and "Theft" topics show small variations in their cardinality along the period from December 2019 to February 2020. It is interesting to see how objects belonging to clusters C3 and C6, which are related to "Waste" and "Theft" topics, respectively, merge in the subsequent window into cluster C10 (related to the "Coronavirus" topic), with a high value of *Preservation*. In fact, cluster C10 groups up both news directly related to the "Coronavirus", and news concerning other topics within the general pandemic context (e.g., how the pandemic affected garbage collection and led to the appearance of wild animals). In the subsequent window, i.e., in April 2020, cluster C10 was further split into four clusters: cluster C7 related to the "Theft" topic, and clusters C11, C12, and C13 related to the "Coronavirus" topic. By inspecting each cluster and its news, it is clear that all the clusters are related to the pandemic, but

Table 8. List of Clusters Identified by an ID with Their Tags and Cosine Similarity Values

ID	Tags
C1	waste emergency: 0.839, waste: 0.809, toxic fires: 0.702
C2	waste emergency: 0.777, waste: 0.741, road accidents: 0.695
C3	waste emergency: 0.733, ring road: 0.696, road accidents: 0.692
C4	thefts: 0.827, robberies: 0.773, car theft: 0.760
C5	thefts: 0.846, robberies: 0.765, car theft: 0.755
C6	thefts: 0.806, robberies: 0.777, car theft: 0.737
C7	thefts: 0.802, robberies: 0.745, car theft: 0.723
C8	coronavirus: 0.951, deaths: 0.530, meningitis: 0.498
C9	coronavirus: 0.989, meningitis: 0.474, mice: 0.414
C10	coronavirus: 0.965, meningitis: 0.488, animal abuse: 0.443
C11	coronavirus: 0.997, meningitis: 0.466, mice: 0.372
C12	coronavirus: 0.998, meningitis: 0.466, mice: 0.365
C13	coronavirus: 0.831, road works: 0.576, ring road: 0.571
C14	coronavirus: 0.996, meningitis: 0.468, mice: 0.381
C15	coronavirus: 0.811, waste emergency: 0.575, animal abuse: 0.566

Table 9. Examples of Headlines Articles for Three "Coronavirus"-Related Clusters in April–May 2020

ID	Headline of articles
C11	Coronavirus: the current positive cases are 4562 but in the last 24 hours there have been 17 deaths.
C11	Coronavirus, 1901 positive cases in Lazio. In Rome the trend of infected people is decreasing.
C12	Pasta and vegetables delivered at home: the brigade bring the shopping home of the most fragile.
C12	Yeast hard to find in supermarkets: the alternatives to make it at home.
C13	Bikes, cycle paths and scooters: the Capitol accelerates on soft mobility to tackle phase 2.
C13	Phase 2, car dealers reopen: who can go there.

each one provides a different perspective. Specifically, cluster C11 contains news about recently detected coronavirus cases and restrictions to mitigate the increase of these new cases, whereas in C12, the articles describe the new delivery ways offered by restaurants and supermarkets for their products, and some services offered to doctors and nurses; cluster C13 deals with all the municipality works on facilities, and improvements to roads to promote the use of bikes, scooters, or eco-friendly ways of commuting. Some examples of these news articles are reported in Table 9.

Notably, only the labeling of cluster C13 precisely matches the actual content, with tags like "road works" and "ring road" that relate to the semantic area of urban mobility. In C11 and C12, instead, the secondary tags ("mice" and "meningitis") are rather distant from the cluster centroid tags in the embedding space (cosine similarity lower than 0.5). This highlights the importance of using an accurate reference tag dictionary, which should cover a range of semantic areas as wide as possible: The current set, for example, does not include terms in the field of food and restaurants, and this likely motivates the apparently whimsical labeling of cluster C12.

Finally, we observe an occurrence of the *Topic Fusion* pattern in the last window (May 2020), involving clusters C11, C12, and C14.

7.6 GeoSpatial Analysis

The proposed approach can take advantage of geo-localization of news, and can help in getting to a more focused analysis by considering specific areas of the city, and visualize the evolution of

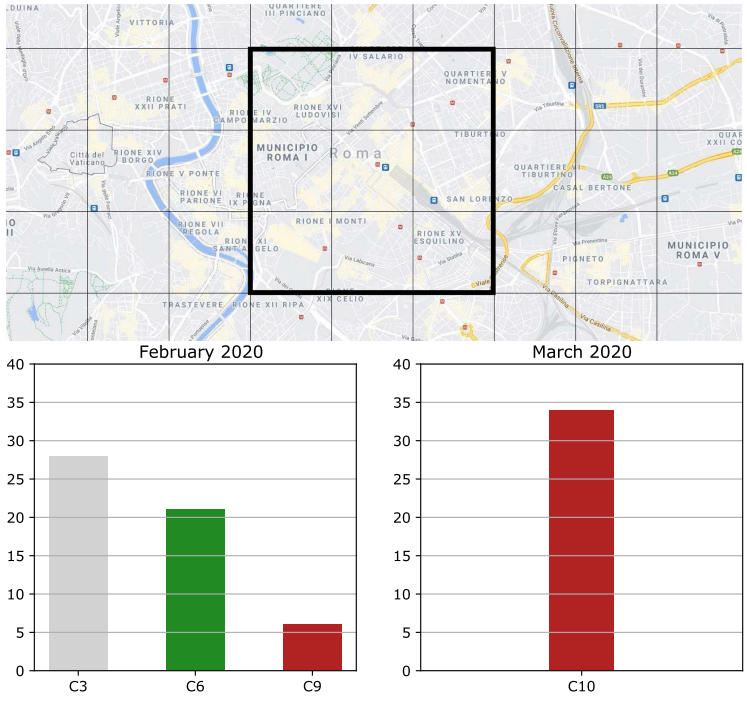


Fig. 9. Evolution of clusters in a specific area near the Roma Termini train station along two subsequent time windows. Each cluster is identified by an ID; see Table 8 for the relative tags.

the news pertinent to such areas. Here, we provide an example of the approach exploited to visualize the clusters of news for the same part of the city in two different periods of time. We chose the area around the Termini Railway Station (the biggest train station in Rome), which is densely populated and was deeply affected by the pandemic. We visualized the news clusters for two time windows, one directly before the enforcement of the Nation-wide lockdown, and the other during the lockdown itself. Figure 9 shows the relative results: most of the news concerning the city area before the major lockdown, represented by clusters C6 and C3, refer to small crimes (e.g., thefts, robberies, car thefts), which unfortunately often affect areas of this kind in major cities, and to waste problems and road accidents, respectively. Cluster C9 shows up as a minor structure linked to the Coronavirus pandemic likely due to the initial restrictions enforced in February. Conversely, in the next period, i.e., during the first Nation-wide lockdown, reports of robbery and waste emergency suddenly disappear, and most of the news articles are targeted toward the pandemic.

Whereas previous experiments allowed us to evaluate the method as a whole, narrower analyses like the one reported in this section may help provide a clearer view of how events unfold in specific city areas during specific periods of time.

8 CONCLUSIONS AND FUTURE WORKS

In this work, we have proposed an approach that is able, on the basis of an analysis of online news, to monitor and track changes in daily life in different areas of the city, both in normal conditions and in the case of extraordinary events, where changes may be more noticeable and bear more complex consequences. The pipeline used in the supporting framework includes: (i) a Sentence-BERT pre-trained language model to obtain news article-level representations

that are semantically relevant, and can be compared in terms of proximity by cosine similarity; (ii) a modified version of the TSF-DBSCAN clustering algorithm for grouping up articles that cover similar topics; (iii) a method for the automatic labeling of the identified clusters; and (iv) a set of metrics aimed at relating clusters generated in consecutive windows.

We evaluated our approach through an extensive case study on the city of Rome during the Covid-19 pandemic, and specifically over the first Nation-wide lockdown. The experimentation performed with the deployment of our news-based framework and the related methodology for knowledge extraction from discovered clusters can give us a few interesting insights.

First, we have observed that the methodology is effective in isolating the specificity of the pandemic, and the consequent lockdown, by means of the clusters of news in specific moments in time and in specific areas. We also showed how a single neighborhood of the city, and particularly one that was deeply affected by the lockdown, drastically changed in terms of news reported for the area.

Second, we adapted a fuzzy density-based clustering algorithm for data streams, i.e., TSF-DBSCAN, to manage high-dimensional data and to automatically tune its parameters from the actual density of objects. The Dimensionality Reduction sub-module supports the new version of TSF-DBSCAN (i.e., *TSF-DBSCAN.news*) to properly deal with word and sequence embeddings produced with state-of-the-art Language Models, characterized by a high dimensionality. We have shown that the UMAP algorithm is an effective choice for reducing the number of dimensions of distributed representations of texts, thus making *TSF-DBSCAN.news* a viable solution for high-dimensional data and NLP applications. Furthermore, the automatic tuning enables *TSF-DBSCAN.news* to autonomously manage streams of data with possibly changing distributions.

Third, the metrics purposely introduced in this work for monitoring the evolution of clusters over time, especially in this context, help analyze how the whole partition, as well as any single clusters, change over time. The three metrics, namely Purity, Coverage, and Preservation, used either singularly or in conjunction with one another, allow uncovering evolutionary relationships across clusters in adjacent time windows, thus providing the ability to track how the profiles of the city areas change along the time.

Finally, it is worth underlining a major advantage of our approach: no supervised models are used in the overall pipeline; this allows us to avoid the costly and time-consuming labeling of single pieces of news as pertaining to certain topics. In the future, we plan to extend the proposed framework to make it operate in distributed environments (e.g., with nodes dedicated to specific tasks and/or specific cities). Moreover, we aim at studying how to improve the labeling technique as well, in order to automatically identify highly informative tags across time. In addition to this, the obtained results could pave the way to proficiently support different kinds of applications that may benefit from the outcome of city area profiling, thus enabling the integration of an additional layer of information.

In conclusion, the proposed method can be an effective tool for monitoring changes in the life of a given city by the plain use of newspaper information readily available on the web. Moreover, the framework turns to be effective also in understanding how changes take place during crisis times like the global Covid-19 pandemic, how such changes impact the life in the city, and it potentially provides insights to organizations, either public ones or others, to better cope with such hard times and to properly and effectively react.

REFERENCES

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. 2014. Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of the 2014 International Conference on Multimodal Interaction*. 427–434. DOI: https://doi.org/10.1145/2663204.2663254

[2] A. Calafiore, G. Palmer, S. Comber, D. Arribas-Bel, and A. Singleton. 2021. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems* 85 (2021), 101539. DOI: https://doi.org/10.1016/j.compenvurbsys.2020.101539

[3] A. Del Bimbo, A. Ferracani, D. Pezzatini, F. D'Amato, and M. Sereni. 2014. LiveCities: Revealing the pulse of cities by location- based social networks venues and users analysis. In *Proceedings of the 23rd International Conference on World Wide Web*. 163–166. DOI: https://doi.org/10.1145/2567948.2577035

[4] A. Ullah, K. Haydarov, I. U. Haq, K. Muhammad, S. Rho, M. Lee, and S. W. Baik. 2020. Deep learning assisted buildings energy consumption profiling using smart meter data. *Sensors (Switzerland)* 20, 3 (2020), 873. DOI: https://doi.org/10.3390/s20030873

[5] Adnan Abid, Ansar Abbas, Adel Khelifi, Muhammad Shoaib Farooq, Razi Iqbal, and Uzma Farooq. 2020. An architectural framework for information integration using machine learning approaches for smart city security profiling. *International Journal of Distributed Sensor Networks* 16, 10 (2020). DOI: https://doi.org/10.1177/1550147720965473

[6] James Allan, Jaime G. Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 194–218.

[7] Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. arXiv e-prints, Article arXiv:2008.09470 (Aug. 2020), arXiv:2008.09470 pages. arXiv:2008.09470 [cs.CL].

[8] Nafis Anwari, Md. Tawkir Ahmed, Md. Rakibul Islam, Md. Hadiuzzaman, and Shohel Amin. 2021. Exploring the travel behavior changes caused by the COVID-19 crisis: A case study for a developing country. *Transportation Research Interdisciplinary Perspectives* 9 (2021), 100334. DOI: https://doi.org/10.1016/j.trip.2021.100334

[9] Joel Azzopardi and Christopher Staff. 2012. Incremental clustering of news reports. *Algorithms* 5, 3 (2012), 364–378. DOI: https://doi.org/10.3390/a5030364

[10] Alessio Bechini, Martina Criscione, Pietro Ducange, Francesco Marcelloni, and Alessandro Renda. 2020. FDBSCAN-APT: A fuzzy density-based clustering algorithm with automatic parameter tuning. In *Proceedings of the 2020 IEEE International Conference on Fuzzy Systems*. 1–8. DOI: https://doi.org/10.1109/FUZZ48607.2020.9177702

[11] Alessio Bechini, Francesco Marcelloni, and Alessandro Renda. 2022. TSF-DBSCAN: A novel fuzzy density-based approach for clustering unbounded data streams. *IEEE Transactions on Fuzzy Systems* 30, 3 (2022), 623–637. DOI: https://doi.org/10.1109/TFUZZ.2020.3042645

[12] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When is "nearest neighbor" meaningful? In *Proceedings of the International Conference on Database Theory*, Catriel Beeri and Peter Buneman (Eds.). Springer, Berlin, 217–235.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (06 2017), 135–146. DOI: https://doi.org/10.1162/tacl_a_00051

[14] Giovanni Bonaccorsi, Francesco Pierri, Matteo Cinelli, Andrea Flori, Alessandro Galeazzi, Francesco Porcelli, Ana Lucia Schmidt, Carlo Michele Valensise, Antonio Scala, Walter Quattrociocchi, and Fabio Pammolli. 2020. Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15530–15535. DOI: https://doi.org/10.1073/pnas.2007658117

[15] Alessandro Bondielli, Pietro Ducange, and Francesco Marcelloni. 2020. Exploiting categorization of online news for profiling city areas. In *Proceedings of the 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems*. 1–8. DOI: https://doi.org/10.1109/EAIS48028.2020.9122777

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, 632–642. DOI: https://doi.org/10.18653/v1/d15-1075

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. Retrieved from https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] C. Zhong, S. Zeng, W. Tu, and M. Yoshida. 2018. Profiling the spatial structure of London: From individual tweets to aggregated functional zones. *ISPRS International Journal of Geo-Information* 7, 10 (2018), 386. DOI: https://doi.org/10.3390/ijgi7100386

[19] Livio Cascone, Pietro Ducange, and Francesco Marcelloni. 2019. Exploiting online newspaper articles metadata for profiling city areas. In *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*,

Hujun Yin, David Camacho, Peter Tino, Antonio J. Tallón-Ballesteros, Ronaldo Menezes, and Richard Allmendinger (Eds.). Springer International Publishing, Cham, 203–215.

[20] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. ACL, 1–14. DOI: https://doi.org/10.18653/v1/S17-2001

[21] Rafael Christófano, Wilson Marcílio Júnior, and Danilo Eler. 2021. PlaceProfile: Employing visual and cluster analysis to profile regions based on points of interest. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1*. INSTICC, SciTePress, 506–514. DOI: https://doi.org/10.5220/0010453405060514

[22] Eleonora D'Andrea, Pietro Ducange, Danilo Loffreno, Francesco Marcelloni, and Tommaso Zaccone. 2018. Smart profiling of city areas based on web data. In *Proceedings of the 2018 IEEE International Conference on Smart Computing*. 226–233. DOI: https://doi.org/10.1109/SMARTCOMP.2018.00070

[23] Eleonora D'Andrea and Francesco Marcelloni. 2017. Detection of traffic congestion and incidents from GPS trace analysis. *Expert Systems with Applications* 73 (2017), 43–56. DOI: https://doi.org/10.1016/j.eswa.2016.12.018

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. ACL, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[25] E. Kaufman. 2016. Policing mobilities through bio-spatial profiling in New York City. *Political Geography* 55 (2016), 72–81. DOI: https://doi.org/10.1016/j.polgeo.2016.07.006

[26] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 226–231.

[27] Anthony Evans, Plamen Angelov, and Xiaowei Zhou. 2006. Online evolving clustering of web documents. In *Proceedings of the 2nd Annual Symposium on Nature Inspired Smart Adaptive Systems*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.5129&rep=rep1&type=pdf.

[28] Deborah Falcone, Cecilia Mascolo, Carmela Comito, Domenico Talia, and Jon Crowcroft. 2014. What is this place? Inferring place categories through user patterns identification in geo-tagged tweets. In *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*. 10–19. DOI: https://doi.org/10.4108/icst.mobicase.2014.257683

[29] João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys* 46, 4, Article 44 (Mar 2014), 37 pages. DOI: https://doi.org/10.1145/2523813

[30] Rong Gao, Jing Li, Xuefei Li, Chengfang Song, and Yifei Zhou. 2018. A personalized point-of-interest recommendation model via fusion of geo-social information. *Neurocomputing* 273, C (2018), 159–170. DOI: https://doi.org/10.1016/j.neucom.2017.08.020

[31] Maria Giatsoglou, Despoina Chatzakou, Vasiliki Gkatziaki, Athena Vakali, and Leonidas Anthopoulos. 2016. CityPulse: A platform prototype for smart city social data mining. *Journal of the Knowledge Economy* 7, 2 (1 Jun 2016), 344–372. DOI: https://doi.org/10.1007/s13132-016-0370-z

[32] H. Hassan, S. Shohaimi, and N. R. Hashim. 2013. Risk mapping of dengue in Selangor and Kuala Lumpur, Malaysia. *Geospatial Health* 7, 1 (2013), 21–25. DOI: https://doi.org/10.4081/gh.2012.101

[33] Zellig S. Harris. 1954. Distributional structure. *Word* 10, 2–3 (1954), 146–162. DOI: https://doi.org/10.1080/00437956.1954.11659520

[34] J. Bejar, S. Alvarez, D. Garcia, I. Gomez, L.Oliva, A. Tejeda, and J. Vazquez-Salceda. 2016. Discovery of spatio-temporal patterns from location-based social networks. *Journal of Experimental and Theoretical Artificial Intelligence* 28, 1–2 (2016), 313–329. DOI: https://doi.org/10.1080/0952813X.2015.1024492

[35] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. ACL, 427–431. Retrieved from https://aclanthology.org/E17-2068.

[36] Dmitry Kobak and George C. Linderman. 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology* 39, 2 (1 Feb 2021), 156–157. DOI: https://doi.org/10.1038/s41587-020-00809-z

[37] András Komáromy and Paras Mehta. 2018. LocXplore: A system for profiling urban regions. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, 568–571. DOI: https://doi.org/10.1145/3274895.3274924

[38] Andreas Komninos, Jeries Besharat, Denzil Ferreira, and John Garofalakis. 2013. HotCity: Enhancing ubiquitous maps with social context heatmaps. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*. ACM, New York, NY, Article 52, 10 pages. DOI: https://doi.org/10.1145/2541831.2543694

[39] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data* 3, 1, Article 1 (Mar 2009), 58 pages. DOI: https://doi.org/10.1145/1497577.1497578

[40] L. Salvati, A. Ferrara, and F. Chelli. 2018. Long-term growth and metropolitan spatial structures: An analysis of factors influencing urban patch size under different economic cycles. *Geografisk Tidsskrift - Danish Journal of Geography* 118, 1 (2018), 56–71. DOI: https://doi.org/10.1080/00167223.2017.1386582

[41] George C. Linderman and Stefan Steinerberger. 2019. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science* 1, 2 (2019), 313–332. DOI: https://doi.org/10.1137/18M1216134

[42] M. Sajjad, J. C. L. Chan, and S. S. Chopra. 2021. Rethinking disaster resilience in high-density cities: Towards an urban resilience knowledge system. *Sustainable Cities and Society* 69 (2021), 102850. DOI: https://doi.org/10.1016/j.scs.2021.102850

[43] Leland McInnes, John Healy, and Steve Astels. 2017. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205. DOI: https://doi.org/10.21105/joss.00205

[44] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software* 3, 29 (2018), 861. DOI: https://doi.org/10.21105/joss.00861

[45] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*. Retrieved from http://arxiv.org/abs/1301.3781.

[46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Curran Associates Inc., Red Hook, NY, 3111–3119.

[47] Maitri P. Naik, Harshadkumar B. Prajapati, and Vipul K. Dabhi. 2015. A survey on semantic document clustering. In *Proceedings of the 2015 IEEE International Conference on Electrical, Computer and Communication Technologies*. 1–10. DOI: https://doi.org/10.1109/ICECCT.2015.7226036

[48] Kanika Narang, Seema Nagar, Sameep Mehta, L. V. Subramaniam, and Kuntal Dey. 2013. Discovery and analysis of evolving topical social discussions on unstructured microblogs. In *Advances in Information Retrieval*, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.). Springer, Berlin, 545–556. DOI: https://doi.org/10.1007/978-3-642-36973-5_46

[49] Amy E. Nivette, Renee Zahnow, Raul Aguilar, Andri Ahven, Shai Amram, Barak Ariel, María José Arosemena Burbano, Roberta Astolfi, Dirk Baier, Hyung-Min Bark, Joris E. H. Beijers, Marcelo Bergman, Gregory Breetzke, I. Alberto Concha-Eastman, Sophie Curtis-Ham, Ryan Davenport, Carlos Díaz, Diego Fleitas, Manne Gerell, Kwang-Ho Jang, Juha Kääriäinen, Tapio Lappi-Seppälä, Woon-Sik Lim, Rosa Loureiro Revilla, Lorraine Mazerolle, Gorazd Meško, Noemí Pereda, Maria F. T. Peres, Rubén Poblete-Cazenave, Simon Rose, Robert Svensson, Nico Trajtenberg, Tanja van der Lippe, Joran Veldkamp, Carlos J. Vilalta Perdomo, and Manuel P. Eisner. 2021. A global analysis of the impact of COVID-19 stay-at-home restrictions on crime. *Nature Human Behaviour* 5, 7 (1 Jul 2021), 868–877. DOI: https://doi.org/10.1038/s41562-021-01139-z

[50] Anastasios Noulas, Cecilia Mascolo, and Enrique Frias-Martinez. 2013. Exploiting foursquare and cellular data to infer user activity in urban environments. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management*, Vol. 1. 167–176. DOI: https://doi.org/10.1109/MDM.2013.27

[51] R. C. Estoque, M. Ooba, X. T. Seposo, T. Togawa, Y. Hijioka, K. Takahashi, and S. Nakamura. 2020. Heat health risk assessment in philippine cities using remotely sensed data and social-ecological indicators. *Nature Communications* 11, 1 (2020), 1–12. DOI: https://doi.org/10.1038/s41467-020-15218-8

[52] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. ACL, 3982–3992. DOI: https://doi.org/10.18653/v1/D19-1410

[53] Gilberto Rivera, Rogelio Florencia, Vicente García, Alejandro Ruiz, and J. Patricia Sánchez-Solís. 2020. News classification for identifying traffic incident points in a Spanish-speaking country: A real-world case study of class imbalance learning. *Applied Sciences* 10, 18 (2020), 6253. DOI: https://doi.org/10.3390/app10186253

[54] Kailash Karthik Saravanakumar, Miguel Ballesteros, Muthu Kumar Chandrasekaran, and Kathleen McKeown. 2021. Event-driven news stream clustering using entity-aware contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. ACL, 2330–2340. DOI: https://doi.org/10.18653/v1/2021.eacl-main.198

[55] Samuel Asumadu Sarkodie and Phebe Asantewaa Owusu. 2021. Impact of COVID-19 pandemic on waste management. *Environment, Development and Sustainability* 23, 5 (1 May 2021), 7951–7960. DOI: https://doi.org/10.1007/s10668-020-00956-y

[56] S. E. Lane, J. F. Barlow, and C. R. Wood. 2013. An assessment of a three-beam doppler lidar wind profiling method for use in urban areas. *Journal of Wind Engineering and Industrial Aerodynamics* 119 (2013), 53–59. DOI: https://doi.org/10.1016/j.jweia.2013.05.010

[57] Bhagya Nathali Silva, Murad Khan, and Kijun Han. 2018. Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. *Sustainable Cities and Society* 38 (2018), 697–713. DOI: https://doi.org/10.1016/j.scs.2018.01.053

[58] Vivek Kumar Singh, Nisha Tiwari, and Shekhar Garg. 2011. Document clustering using k-means, heuristic k-means and fuzzy c-means. In *Proceedings of the 2011 International Conference on Computational Intelligence and Communication Networks*. 297–301. DOI: https://doi.org/10.1109/CICN.2011.62

[59] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. Retrieved from http://jmlr.org/papers/v9/vandermaaten08a.html.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[61] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. ACL, 1112–1122. DOI: https://doi.org/10.18653/v1/N18-1101

[62] Chuanjie Yang, Guofeng Su, and Jianguo Chen. 2017. Using big data to enhance crisis response and disaster resilience for a smart city. In *Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis*. 504–507. DOI: https://doi.org/10.1109/ICBDA.2017.8078684

[63] Leyla Zhuhadar, Evelyn Thrasher, Scarlett Marklin, and Patricia Ordóñez de Pablos. 2017. The next wave of innovation-review of smart cities intelligent operation systems. *Computers in Human Behavior* 66, C (2017), 273–281. DOI: https://doi.org/10.1016/j.chb.2016.09.030

[64] Alaettin Zubaroğlu and Volkan Atalay. 2021. Data stream clustering: A review. *Artificial Intelligence Review* 54, 2 (1 Feb 2021), 1201–1236. DOI: https://doi.org/10.1007/s10462-020-09874-x

Received July 2021; revised March 2022; accepted April 2022