future internet

MDPI

Article

A Semantic Preprocessing Framework for Breaking News Detection to Support Future Drone Journalism Services

Michail Niarchos *, Marina Eirini Stamatiadou, Charalampos Dimoulas, Andreas Veglis and Andreas Symeonidis

School of Journalism & Mass Communication, Aristotle University of Thessaloniki, 54636 Thessaloniki, Greece; mstamat@jour.auth.gr (M.E.S.); babis@eng.auth.gr (C.D.); veglis@jour.auth.gr (A.V.); asymeon@eng.auth.gr (A.S.)
* Correspondence: mniarchos@jour.auth.gr

Abstract: Nowadays, news coverage implies the existence of video footage and sound, from which arises the need for fast reflexes by media organizations. Social media and mobile journalists assist in fulfilling this requirement, but quick on-site presence is not always feasible. In the past few years, Unmanned Aerial Vehicles (UAVs), and specifically drones, have evolved to accessible recreational and business tools. Drones could help journalists and news organizations capture and share breaking news stories. Media corporations and individual professionals are waiting for the appropriate flight regulation and data handling framework to enable their usage to become widespread. Drone journalism services upgrade the usage of drones in day-to-day news reporting operations, offering multiple benefits. This paper proposes a system for operating an individual drone or a set of drones, aiming to mediate real-time breaking news coverage. Apart from the definition of the system requirements and the architecture design of the whole system, the current work focuses on data retrieval and the semantics preprocessing framework that will be the basis of the final implementation. The ultimate goal of this project is to implement a whole system that will utilize data retrieved from news media organizations, social media, and mobile journalists to provide alerts, geolocation inference, and flight planning.

Keywords: breaking news; semantic processing; natural language processing (NLP); drone journalism; events location estimation

Citation: Niarchos, M.; Stamatiadou, M.E.; Dimoulas, C.; Veglis, A.; Symeonidis, A. A Semantic Preprocessing Framework for Breaking News Detection to Support Future Drone Journalism Services. *Future Internet* **2022**, *14*, 26. https://doi.org/10.3390/fi14010026

Academic Editor: Luis Javier Garcia Villalba

Received: 10 November 2021
Accepted: 6 January 2022
Published: 10 January 2022

1. Introduction

When it comes to event and breaking news coverage, textual representation can be supplemented with live audio–visual (AV) footage to ensure objective and reliable news reporting. While citizen and mobile journalists can offer real-time coverage through social media and User-Generated Content (UGC), they are not always capable of providing quality footage or even accessing the location of interest. On top of this, in the era of the COVID-19 pandemic, mobility is usually restricted, meaning that entering a once easily accessible place can become difficult or even impossible for a person. During the last decade, drones have democratized landscape aerial photography, making it affordable even for individual professionals and freelancers as well. Hiring or purchasing expensive equipment, such as helicopters, is no longer needed. At the same time, drones can be used to access areas that journalists would prefer not to visit while collecting high-quality AV content using their constantly improving (camera/microphone) equipment. Their relatively small size also allows their fast and effective deployment in many cases.

Moving from plain text towards AV news coverage and storytelling will improve the communication of a story [1,2]. AV data are not only valuable for the news consumers, but these data can be proven to be useful for other professionals during post-processing. This introduces the demand for high-quality footage for the data to be efficiently used by image processing tools [3–6].

Even though it is undoubted that drones can be used in several on-field scenarios, their usage is not yet commonplace. The main reason for this is that ethics and safety concerns have led to the adoption of strict operating policies in most countries [7,8]. In addition, problems concerning intrusiveness and aesthetic ideals have been posed by photojournalists [9]. The future relaxation of the commercial application restrictions will enable the media industry and individual reporters/professionals to consistently use them. In the United States, the New York Times and the Washington Post are already using drones to take photos and videos of simulated news events at an FAA-designated test site at Virginia Tech University [10]. They have also started preparing the new generation of "dronalists", i.e., journalists trained as Pilots in Command (PICs), to use this new, exciting technology in day-to-day operations. In this spirit, this work proposes a framework for future drone journalism services, which will be fully deployable as soon as regulations have been conducted to overcome the aforementioned restrictions.

The usage of computational and robotic resources is becoming widespread in the ecosystem of Journalism 3.0. A variety of emerging journalistic forms (textual, audio, and pictorial) are enriching online media content [11]. Natural language generation (NLG) is changing production, as it can perform functions of professional journalism on a technical level [12], leading to automatic news generation [13], sometimes making use of Machine Learning (ML) and/or natural language processing (NLP) [14–17]. ML is defined as the study of algorithms that use computer systems to perform tasks without being instructed to (i.e., learning by example), while NLP, which is a part of Artificial Intelligence (AI), deals with the analysis and processing of human natural language [18]. While multi-source news retrieval is becoming commonplace, methods are in development to verify the newsgathering and the publication process [19–21]. Dedicated search and retrieval mechanisms, implemented as software agents, can be applied in data that are stored in intermediate storage locations, inferring useful information for further analysis and monitoring [22]. Even with the absence of prior knowledge, it is possible to extract events based on semantics in text and/or speech [23] or a specific timeline of events [24].

Paving the way for "dronalism", the need for the automated deployment of drones as autonomous unmanned robotic journalists has emerged. The process of news retrieval and real-time coverage of an event can be highly automated. Our hypothesis is that today's technology is mature enough to allow the implementation of a service for automated breaking news detection and notification-based communication with a drone management subsystem. The main goal is the enhancement of journalistic workflows in news coverage. More specifically, this paper aims at launching prototype automation services in drone journalism, facilitating event detection and spatiotemporal localization processes until the flight-plan execution and the news-scene coverage completion. Prototype data monitoring and recording processes will be deployed (retrieval, normalization, processing, and warehousing) to organize, classify, and annotate content/metadata properly. Stream queries and analysis will rely on the articles/posts of news-sites and social networks (blogs, Twitter, etc.), along with the transcription and indexing of press media, selected radio, and TV broadcasting programs. The aim is to accelerate human processing with machine automation that executes multiple/parallel time-, location-, and context-aware correlations in real time for event-alerting purposes. The validation of the extracted alerts and the exploitation of geographical/map information allows ground access areas to be found and recommended, where Unmanned Aerial Vehicles (UAVs) can take off towards a region of interest. Thereafter, flight-plan preparation and guided/semi-automated flight navigation and coverage can be further investigated, as long as specific criteria are met. These processes make a good fit for this current Special Issue.

One of the key elements that had to be answered from the beginning of the project is the degree to which the potential users would find the envisioned service useful, and to investigate the associated requirements and the preferred functionalities to implement, which is aligned with the analysis phase of standard software development procedures [25,26]. To achieve this, a survey was carefully designed and conducted to serve the needs of audience

analysis. In this context, it is equally essential to implement and validate the applicability of the algorithmic breaking news detection back-end, which stands as the core element for serving the whole idea. Hence, the whole concept had to be tested both in terms of users' acceptance and technologic solutions' applicability. In this direction, NLP systems were implemented as the initial algorithmic solutions and were thoroughly evaluated at various levels to provide a convincing proof of concept of the tested scenario.

The current work proposes a framework for functioning and supporting drone journalism services, incorporating the existing regulatory framework, security, and ethical matters [8,27,28]. In this context, the technological and functional capacities of drones, i.e., the capability to create powerful narratives, will be combined with a structured and automated way of operation, including alerting, control data transfer, and normalization services. The system is modular, able to integrate UGC and news alerts sent by mobile journalist devices, making use of the inherent localization and networking capabilities [25,29–33], or even cooperative drone monitoring, thus offering time-, location- and context-aware data-driven storytelling. This will lead to the faster and more reliable detection of breaking news with their contextual and spatiotemporal attributes, using crowd computer intelligence techniques with appropriate weighting mechanisms. Modularity does not only concern the retrieval and the processing stage but the output forwarding as well. Alerts will also be sent to mobile journalists to let them cover the event too. Diverse drone footage and UGC mobile streams can be combined to offer a more holistic coverage of the events.

The designed system is based on three pillars:

1. Breaking news detection and geolocation inference, which produce notifications to be sent to a drone handling system or individual drone operators.
2. Flight-plan preparation.
3. Semi-autonomous waypoint-based guidance.

Focus is given on data retrieval and, mainly in the preprocessing phase of the first stage, the whole system is built adopting a modular design approach, consisting of independent subsystems, the implementation of which is outside of this paper's focus and context.

The rest of the paper is organized as follows. A literature review around breaking news detection is presented in Section 2. Section 3 describes the assumptions made while formulating the list of actions needed, based on the opinion of field experts. The system architecture is also described in this section, along with the main functions and the components comprising the developed framework. In Section 4, we present the results of the questionnaire that was distributed for the assessment of the envisioned system, as well as the performance and the evaluation of the breaking news detection method. Conclusions are drawn and a discussion is presented in Section 5.

2. Related Work

Drones are already being used in video capturing and news coverage. Research has been carried out for flight planning and approaching inaccessible locations [34,35]. Such topics are out of the context of this work, as they refer to problems that will be faced in the last chain link of the complete system. The current stage of implementation focuses on data preprocessing and more specifically on breaking news detection; thus, the literature that is presented below refers to such semantic analysis.

One of the novelties and research contributions of the current work lies in the implemented solutions for the specific problem of multi-source breaking news detection, which makes a strong proof of concept of the proposed scenario. Based on the conducted literature/state-of-research review (and to the best of our knowledge), it seems that previous works dealing with the same problem and under similar constraints are very limited and/or entirely missing. The majority of similar approaches process social media messages (mostly tweets) [36–39], and only a few of them exploit website metadata (e.g., HTML tags [40]) or use the whole article text [41] to detect hot words. Most of the recent works seem to deal with hot topic detection, which can also be adapted to breaking news detection. However, these are two different problems in principle, i.e., "breaking news" implies

urgency, while "hot topic" implies popularity. Overall, the current work proposes a holistic treatment incorporating the detection of breaking news from unstructured data emanated by multiple sources (therefore including potential UGC contributions). The implicated spatiotemporal and contextual awareness perspectives model a semantic processing framework with modular architecture and significant future extensions on drone journalism news covering automations. A common technique for detecting hot topics is the use of keywords [36,41]. Such techniques usually show low precision, as they are based on keyword occurrence frequency in documents written by the general public [36]. The careful selection of keywords can lead to reliable results. The collection of keywords by social media posts can be very efficient. If this collection is extracted by selected reliable users, taking into consideration their influence and expertise can further improve this method [36]. Bok et al. [36] proposed the use of a modified version of TF-IDF that incorporates a temporal factor to be able to tackle the difficulty of detecting near-future breaking news.

Natural Language Inference (NLI) includes methods for predicting whether the meaning of a piece of text can be inferred by another [42]. A form of NLI is paraphrasing, which is also called text pair comparison. Computing the similarity among pieces of text is used in various applications, and it can be applied in this use case as well. Jensen–Shannon divergence is used to measure the similarity between two topics, which is used in topic tracking [43].

An interesting approach is the TDT_CC algorithm—Hot Topic Detection based on Chain of Causes [37]. This algorithm treats events as topic attributes and uses them to update a graph, aiming to detect a trend in a large graph in real time. Traditional algorithms dedicate too much time to traversing a graph, while TDT_CC tackles this issue by focusing on the structural topology.

Graphs are also used to effectively compute the relevance between textual excerpts [38]. Hoang et al. [38] utilized a term graph to measure the relevance between a tweet and a given breaking event.

Clustering is used in all of the above methods at some stage. Shukla et al. [39] applied clustering to filtered, collected tweets and then scored the clusters based on the number of tweets, which led them to breaking news detection. This is a simple and reliable technique which can only be applied on social media. In addition, this technique is not efficient in terms of memory usage and execution time, which makes it quite unusable in real-world deployments.

Applications that crawl specific websites to gather news can utilize various forms of metadata to enhance their detection methods. HTML tags constitute such a kind of metadata. They can be used to isolate the title, the subtitle, or any other semantically important piece of information to achieve a better detection rate [40]. The major defect of such an approach is that it is bound to the structure of certain platforms and is vulnerable to any possible structural change.

3. Materials and Methods

3.1. Assumptions and Setup

In order to validate our concept hypothesis regarding its user acceptance, an empirical survey was conducted in the form of a questionnaire distributed online to 100 people. Data were collected within one month (February–March 2021). Typical questions concerning breaking news capturing, reporting, and sharing were answered, retrieving vital feedback. Hence, background- and general-interest-related questions were structured in a categorical form of potential answers, with 5-point Likert scales (1–5, from "Totally Disagree" to "Totally Agree"). Binary values (i.e., gender) and higher-dimension lists were also involved. The items were divided into three subsets, with the former involving questions regarding the current state of/trends in (breaking) news reporting (Q1–Q4), the second implicating questions on the envisaged modalities and usability characteristics of the proposed service (Q5–Q9), and the latter containing basic characteristics/demographics of the users (Q10–Q17). The survey formation was validated after discussions and focus groups with

representative users. Specifically, both professional and citizen journalists were involved, as well as people from the broader journalistic and news reporting field. The survey was updated based on the received feedback, investigating the audience interest in a system that incorporates mechanisms for automatic breaking news detection and their intention to contribute with their content. The gathered information was used for the estimation of the anticipated dynamics of the proposed architectural approach. An overview of the chosen inquiries is presented here, aiming to justify the adoption and configuration of the formed questionnaire. Detailed information regarding this survey is provided in the associated Results section, along with the assessment outcomes.

During the survey preparation, all ethical approval procedures and rules suggested by the "Committee on Research Ethics and Conduct" of the Aristotle University of Thessaloniki were followed. The respective guidelines and information are available online at https://www.rc.auth.gr/ed/ (accessed on 9 July 2021). Moreover, the Declaration of Helsinki and the MDPI directions for the case of pure observatory studies were also taken into account. Specifically, the formed questionnaire was fully anonymized, and the potential participants were informed that they agreed to the stated terms upon sending their final answers. All participants had the option of quitting any time without submitting any data to the system.

3.2. Architecture

The envisioned architecture is presented in Figure 1, along with the main functions. It was designed to target the following objectives:

- To gather original data heterogeneous data sources (web, social media);
- To process and analyze input data to form an event classification taxonomy extracting breaking news;
- To create automated alerts to trigger drone applicability;
- To support semi-automated/supervised drone navigation (and tagging) through a set of waypoints (with the notes that a safe landing zone is always needed, predefined, or detected in deployment time, while PIC in the communication range is required at all times);
- To develop user-friendly dashboards for drone control;
- To synchronize aerial footage with ground-captured AV streams (e.g., UGC) and propagate related annotations based on spatiotemporal and semantic metadata (i.e., topic, timestamps, GPS, etc.);
- To investigate AV processing techniques for the detection of points/areas of interest in arbitrary images and video sequences;
- To provide information, training, and support about drone utilization in public based on local regulations and ethical codes;
- To implement semantic processing and management automation for the captured content in batch mode, which could be utilized in combination with the other publishing channels and UGC streams (i.e., for analysis, enhancement, and publishing purposes).

The architectural implementation provides a solid framework, including:

1. A prototype web service with open access to external (third-party) news sources for the detection of breaking news;
2. Pilot software automating aspects of drone-based newsgathering;
3. Content storage and management repository with indexing and retrieval capabilities;
4. On-demand support and training on the new services.

All legislation matters were accounted for during project deployment and testing and the associated training sessions. Aiming to address both breaking news detection and event geographical localization, field journalism experience was also taken into account.

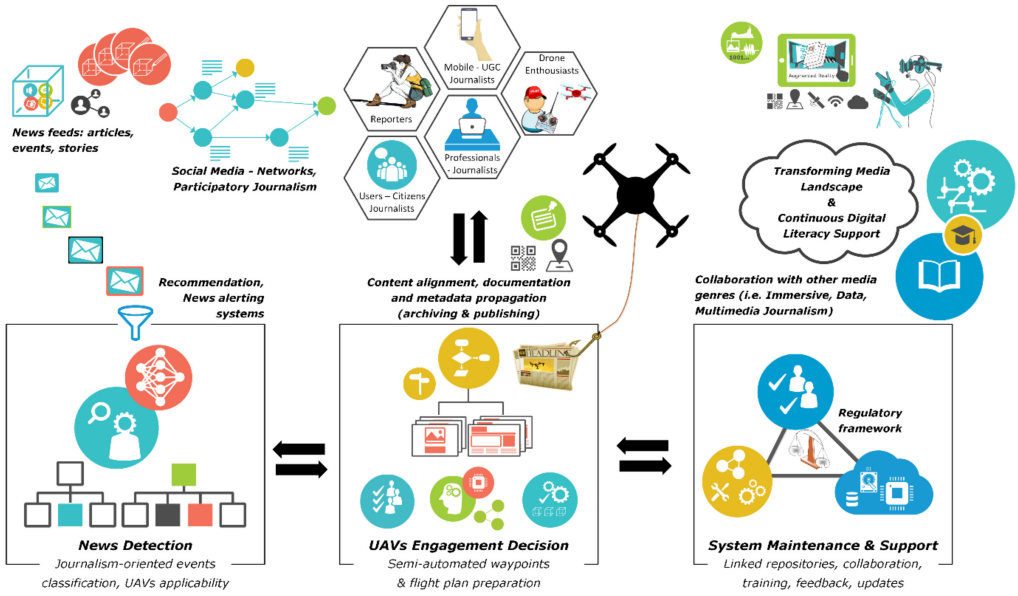


Figure 1. The envisioned concept architecture of the proposed model [8].

3.3. Framework Description

The graphical representation of the developed modular architecture is presented in Figure 2. The underlying components that drive the process of breaking news detection are described in short. Data propagation for automated, real-time drone coverage of breaking news events is presented as well. The concept was evaluated by collaborating closely with a focus group of active journalists that helped to form the outcome.

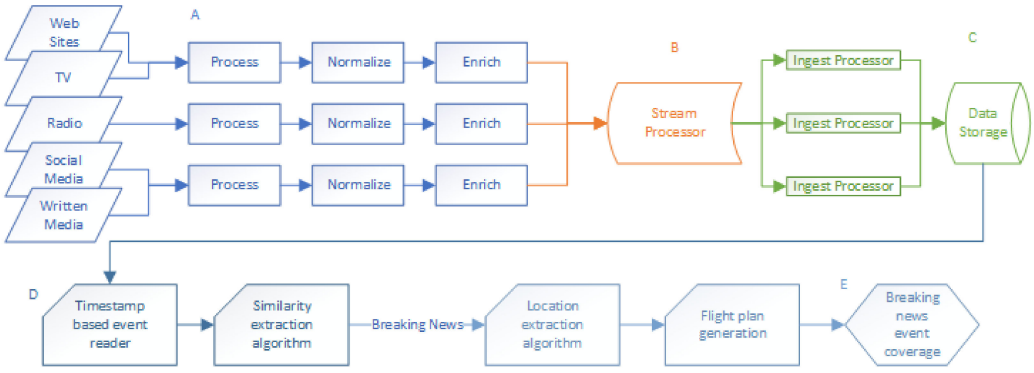


Figure 2. Block diagram of the implemented data flow.

Figure 2 depicts an overview of the implemented data flow. In the architecture, data collection starts upon their creation while monitoring several news feeds for events (A). An extensive effort was put into multi-source event gathering from (a) websites through data scraping and indexing, (b) written press such as newspapers through digitization and indexing, and (c) social media, television, and traditional radio through transcription, which is currently carried out by a news provider collaborator with the use of custom speech-to-text tools. News extracted by TV and radio is considered especially important. Many events may be firstly broadcast by these traditional media, as they still have faster access to specific news agencies. Each event is described as a single and independent data message, generated by a news feed. Such a data message, e.g., an article or a Twitter tweet, may imply one or multiple events, but in terms of news coverage, this can be considered as a single event requiring journalistic action. This initial process produces events to be propagated to the system after being enriched with additional information that allows categorization and management that will subsequently help to effectively query for events.

Events enter the system utilizing a common bus (B), a communication channel that serves the purpose of conveying information while, at the same time, ensuring data integrity and avoiding data redundancy (buffered process). It can store data in an intermediate state for a specified amount of time that can range from minutes to several days. Events that exit the common channel are stored in a database (C) in a structured way that enables fast indexing and retrieval. Events are retrieved (D) as sets of filtered information for specified time windows/ranges and checked for correlation using a multi-step detection algorithm. The output is a number of real events, along with their statistics, marked as breaking news. For a drone to be able to interconnect with such a system and operate, geographical information is needed. This information is extracted using a separate, multi-stage algorithm (E) that takes into account location-related terms in the actual text of the event, as well as related imagery that can be retrieved from it. The system is integrated as a web application that can also accept UGC that enhances user experience.

The framework consists of the modules shown in Figure 3, which outlines the conceptual architectural processes as described above. The innovation of this design lies in the ability of the system to continuously improve its performance, while also being able to handle each different input type independently by inserting new "pluggable" entities into the system (e.g., interaction with the devices of mobile journalists). Detailed information is given in the following subsections.

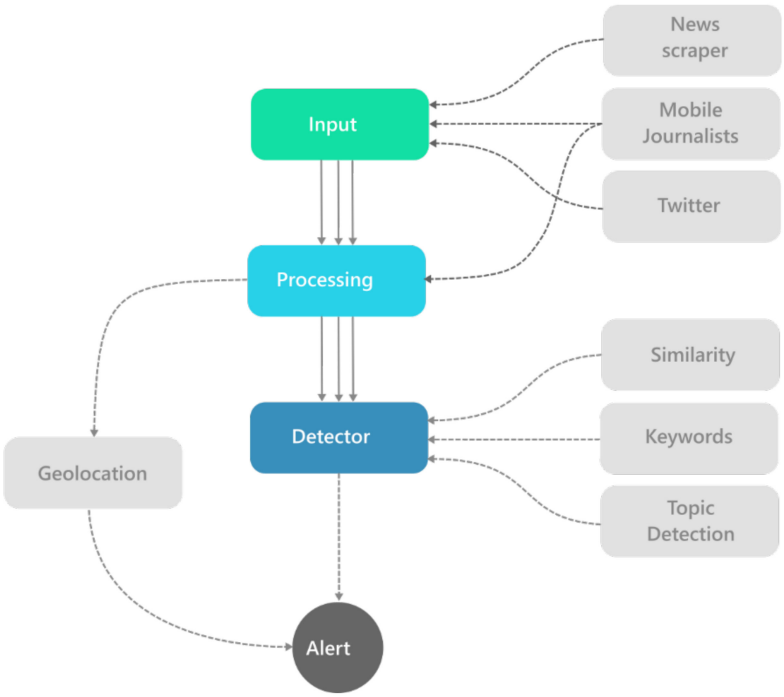


Figure 3. Modularized architecture.

Elaborating on the above, the main aim was to set up and configure a modular architecture, allowing each individual modality (radio, TV, newspaper, etc.) to be added and treated consistently, in the form of text and images. Hence, uniform text and image-derived metadata processing steps are introduced and deployed into an integrated framework. It has to be clarified that individual modules, feeding the framework with input data, are not fully optimized as part of this paper. Specifically, the back-end crawling service of a media monitoring collaborator was utilized and tested, representing a real-world scenario. As already mentioned, the associated utilities rely on OCR and speech-to-text services, supplemented with human intervention. The whole approach aligns with the objectives of the paper to provide an end-to-end pipeline encompassing seamless and uniform treatment of all individual media inputs.

3.4. Data Retrieval and Management

The process of data retrieval is critical since it provides the necessary information for the system to operate, while it introduces serious implementation difficulties. The diversity of available data sources [11] leads to the need for many different ways of handling input, while there is no lack of corner cases that require special treatment. For example, data scraping tools [44] can extract information from websites, but every website has a different structure, which means that each case requires different handling. Moreover, websites may change their presentation layer on a relatively regular basis, trying to make it more appealing to the users. Scraping tools are sensitive to changes at the presentation layer, which leads to the re-engineering of the data collection strategy. The density and the type of information included in the different data sources make the task of classification difficult. Text transcript from a radio broadcast contains less information than a simple post from a social media user profile, yet they need to be stored identically to facilitate processing capability by analytics tools. Information decomposition to title, text, source, etc., is currently executed by the news provider we are collaborating with, but ultimately, a custom implementation will be carried out to support the data retrieval from more sources.

Events are organized as sets of fields, with each one containing specific data types and content. An event schema is presented in Figure 4, along with an example event (translated in English). This process is called normalization and is the cornerstone for the creation of a unified mechanism to execute analytics [45] against really large amounts of information. The final phase taking place is the enrichment process, which can be described as the action of adding artificial fields in the original event that would not normally exist. Internal content and articles' metadata form useful information that feeds the implemented breaking news detection system. This data may include title and relevant keywords, media-related information, and time-, location-, and context-aware metadata, which in terms of NLP research community terminology are transformed to more solid and crisp representations, i.e., action, agents, place, time, etc. Despite the requirements that algorithms introduce, the events are uniformly presented in this form to also be able to propagate the input outside of this internal ecosystem. A crucial criterion behind the schema design is for it to be comprehensible by journalists that will need to process the accompanying information upon breaking news detection.

Column	Description	Type
id	Event unique identifier	integer
title	Article title	string
timestamp	Publication timestamp	string
modification_date	Last modified timestamp	string
medium_id	Medium unique identifier	integer
medium_page_views	Medium page views	integer
medium_text	Medium name	string
snippet	Snippet from the article	string
text	The whole article	string
url	URL to the article	string
is_blog	Identifies if this is a blog post	boolean

(**a**)


```
{
  "id": 440386725,
  "title": "European Parliament ratifies agreement on
           Brexit",
  "timestamp": "2020-01-30 13:28:39",
  "modification_date": "2020-01-30 13:30:51",
  "medium_id": 1108,
  "medium_page_views": 4617285,
  "medium_text": "thesstoday.gr",
  "snippet": "European Parliament votes ... ",
  "text": "European Parliament votes today to ratify
          on the withdrawal treaty and how it will be
          carried out ... ",
  "url: "https://thesstoday.gr/%cf%84%ce%bf-%ce%b5%cf
         %85%cf%81%cf%89%cf%80%ce%b1%cf%8a%ce%ba%cf%8
         c-%ce%ba%ce%bf%ce%b9%ce%bd%ce%bf%ce%b2%ce%bf
         %cf%8d%ce%bb%ce%b9%ce%bf-%ce%b5%cf%80%ce%b9%
         ce%ba%cf%85%cf%81%cf%8e%ce%bd%ce%b5%ce%b9/",
  "is_blog": false
}
```


(**b**)

Figure 4. (**a**) Event schema; (**b**) example event (json).

The output of the previous stage is a set of feeds that generate normalized events. A dedicated stream processor [6,46,47] then handles these data in a parallel manner, ensuring the integrity of the information included. By definition, a stream processor is a queueing system that concurrently accepts input from many sources (publishers) and conveys the events to processes that make use of them (subscribers/consumers). In between, a stream processor implements a buffer that can reliably store events (in transient status) for a specified amount of time (up to several weeks, provided that the required storage resources suffice). It also ensures that events are processed in the same order they entered the system. Under certain conditions, it can also ensure that the events may be replayed in the output at most, at least or exactly once, providing the required integrity. Stream processors can be set up in cluster mode, behaving like a swarm of computational resources that can replicate the information on their buffers in several different physical machines. This behavior

helps to avoid information redundancy usually existing in the buffer, making it possible to recover from software or even hardware breakdowns. Finally, it enables data retrieval by consumers that greatly improves performance through parallelization.

Events come out of the stream processor by a set of consumer processes (ingestors) that are responsible for storing the events in a long-term logbook. The used data format facilitates the storage of massive amounts of data in cloud providers while improving retrieval by sharing data files and optimizing filtered read queries. TileDB [48,49] was chosen as the database system that is responsible for the above procedure. This data engine treats input events as elements in a three-dimensional array, defined as follows: The first dimension is time, expressed in seconds since a specific date and time in the past, mapping to the timestamp field of the event. The second dimension is the actual medium (source) of events. There exists a predefined list of sources identified by special identifiers in integer format. The third and last dimension is the event identifier, which is unique for a specific event and helps in discriminating events that have the same timestamp (measured in seconds). The rest of the available fields of an event are stored in a dedicated cell with coordinates of timestamp, medium, and event as required by the dimensions of the array as attributes. Having the events securely stored using an efficient format allows for their usage in the context of a breaking news reasoner/extractor.

3.5. Breaking News Discovery

Following the setup that was created in Section 3.1, the next step is the utilization of available data that are stored in the medium of our choice. The target is to infer breaking news from the total list of available events. It was decided that a reliable approach would consist of both NLP techniques and expert knowledge; thus, the designed component combines standard computational methods with empirical rules.

Long-time journalists and reporters were engaged to formulate the breaking news detection procedure. Table 1 provides a questionnaire that was initially formed in technical terms (i.e., concerning system implementation) and was further elaborated by incorporating the feedback given by experienced newsroom professionals. While the utmost target is to address these queries to a significant number of media professionals and citizen journalists, an initial validation was obtained by ten (10) targeted users/professionals, different from those involved in the initial formulation process.

The intention was to proceed with an initial parameterization and setup of the proposed system that could be further configured and fine-tuned, upon receiving more audience feedback and testing experience.

The prototype algorithm that was implemented is presented in Algorithm 1. The core of the algorithm is the measurement of the similarity [50–52] of the event title under test with the rest of the titles that have appeared in the near past.

The main advantage of this technique is that it does not require any prior knowledge of the processed text, as it uses a model that has already been trained on a set of documents that serves as a knowledge base. A big data set (500k words) in the Greek language is obtained and then projected to an applicable vector space [53]. Each word in the corpus is represented by a 300-dimensional vector. The vectors are calculated using the standard GloVe algorithm [54], which is a global log-bilinear regression model that combines global matrix factorization and local context window methods. This algorithm is widely used as it performs quite well on similarity tasks. The extracted vectors that are also called word embeddings contain not only the meaning of the word as an arithmetic value but its relation to other words as well. The described process is unsupervised, capable of learning by just analyzing large amounts of texts without user intervention. Moreover, it is possible to use transfer learning—a practice in machine learning where stored knowledge gained while solving one problem can be applied to a different but related problem—on an existing trained data set using a lexicon of general interest to extend its functionality, to achieve a better-contextualized word representation in specific scenarios, where different buzzwords/lexicons are used (e.g., a journalistic dictionary).

Table 1. Empirical rules regarding the criteria for announcing an event as emerging.

Survey Question	Empirical Results, Subject to Test
In your experience, how many common words should the title of two articles contain to be considered to refer to the same event?	One-four or over five words, depending on the length of the title.
In your experience, what is the maximum time difference of the articles under comparison to being considered as referring to the same extraordinary event?	From thirty minutes to over three hours, with the shortest duration to provide stronger indications. However, this does not matter if it is an extraordinary event it will be constantly updated.
When developing an algorithmic system for automated comparison between different articles, would it make sense to use?	Multiple sequential or sliding time-windows, with a degree of overlapping.
What criteria would you use to determine the importance of an article?	Number of sources that simultaneously appear, number of "reliable" sources, specific thematic classification (e.g., natural disaster), number of sharing posts and/or reactions (likes, comments, etc.).
What extra fields would you consider important for the purpose?	Author's name listed, images attached/enclosed, sources and their reliability.

Each title is also represented by a vector, which is calculated by averaging the vectors of the words that constitute it. To measure the relevance between two titles, the cosine similarity between the corresponding vectors is computed [36,39,41,42].

$$S_c(A, B) = \frac{A \times B}{||A||\,||B||} \quad (1)$$

Cosine similarity of zero means that two given sentences have no relevance, while a value of one implies an absolute match. Various experiments were conducted to define the minimum value that implies high contextual relevance. The results of these experiments were given to the field experts that were mentioned earlier, and they set the minimum threshold to the value of 0.85.

Algorithm 1 Breaking news detection

```
while forever do
  ▷ get the articles of the past h hours
  get articles["timestamp >= now-h"]
  for article in articles do
    int counter
    title ← get_title(article)
    title_vector ← get_vec(title)
    for tmp_article in articles do
      tmp_title ← get_title(tmp_article)
        tmp_title_vector ← get_vec(tmp_title)
        s ← similarity(title_vector, tmp_title_article)
        if s >= t then
          increment counter
        end if
      end for
      if counter >= m then
        classify article as breaking
      end if
    end for
end while
```

The detection algorithm is now at the proof of concept stage while it continues to be improved. We are aiming to introduce more modules that will be part of the "detector" super module, each one contributing to the process differently. A topic detection model will also be integrated to improve the filtering of the "candidate" events. The problem of topic detection is not solved, but there have been promising works that introduced automated classification methods [55] that can fit in the proposed framework. On top of this, given that classification algorithms require continuous testing and training using data from reliable databases, focus is given to the continuous enrichment of the designed database, which will help in improving the detection methods.

3.6. Geolocation Inference

The next step is to decide whether the event is eligible for real-time drone coverage or even if a flight plan can be initially designed using waypoints. An important module that will be plugged into the system is the geolocation module. If input data do not contain any GPS tags, then the location has to be algorithmically extracted from the article without compromising the required automation. Combining the knowledge extracted by the group of the ten (10) experts, mentioned in the previous section, and regarding methods mentioned in the recent literature led to the formation of the following list of methods:

- Geo-tags from the associated sources [56];
- Geolocation extraction from named entities, such as [19]:
  - Geopolitical entities (i.e., countries, cities, and states);
  - Locations (i.e., mountains and bodies of water);
  - Faculties (i.e., buildings, airports, highways, etc.);
  - Organizations (i.e., companies, agencies, institutions, etc.).
- Photo documents of the article, through inverse image search/near-duplicate image retrieval [57], followed by visual semantics recognition [58] and geographical metadata propagation.

The above approaches can work in both a fully and semi-automated fashion, expediting the searching process, which can be substantially propelled by human cognition and experience. The incorporation of mobile users (UGC contributions) could assist in the targeted location awareness. In all cases, the system provides an output that optionally includes location information or alerts the user if it is not possible to extract such information.

At the moment, a related base algorithm has been designed and put into a test, making an initial proof of concept. Preliminary results have been extracted, but they still cannot be presented, as the reliability of the implementation is still under dispute.

3.7. Topic Detection

An additional layer of "filtering" is the topic detection/classification module. Topics of specific categories should never appear as breaking news, e.g., lifestyle news. For this reason, NLP and ML methods [59] have been implemented to achieve reliable topic classification and to reduce the input size that will be fed to the detector module. Apart from filtering, this makes it possible to apply weighting based on the topic, as some topics are less likely to be breaking.

In our current approach, the vector space models of the documents are calculated as soon as the documents have been preprocessed (stop-word elimination, stemming, etc.). After this step, clustering is applied based on the cosine similarity of the vectors [60]. This approach may sound simplistic but gives good results when there are only a few topics. The involvement of field experts in the implementation procedure enables the design of a rule-based detection system [42]. Such an approach can be easily implemented from a programming point of view, but it requires deep knowledge of the domain, which makes maintainability difficult. This implementation is still in the tuning phase, as the first conducted evaluation has shown that the percentage of unclassified articles is quite high.

Eventually, deep learning approaches [37] are going to be implemented and compared with the aforementioned one to further evaluate its performance. The main advantage of such methods is that they do not require deep domain knowledge, which leads to faster implantation and makes maintenance and further development much easier. Beyond traditional deep learning techniques, there are novel methods such as Deep Reinforced Learning [21], which can be utilized to achieve even better results.

4. Experimental Results

4.1. Concept Validation through Audience Analysis

To examine audience acceptance of the proposed approach, we undertook an online survey (N = 100). Online surveys allow for generalized data collection, and they are proven to be feasible methods to reach a broad and representative sample. Table 2 synopsizes the final set of questions selected for the needs of this survey. In general, the results showed that many people are not familiar with what DJ is. The majority of the participants expressed their interest in the mediated breaking news experience that is aimed to be achieved within the current project, as thoroughly analyzed below.

Table 2. The analysis questionnaire was answered by 100 people. (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree).

	Questions Answered in Likert Scale	1	2	3	4	5
Q1	I am interested in breaking news	0	4%	15%	40%	41%
Q2	Cross check upon breaking news alert is necessary	0	0	8%	46%	46%
Q3	An automated breaking news alert system would be useful	1%	6%	16%	59%	18%
Q4	I would use an automated breaking news alert system	2%	4%	19%	53%	22%
Q5	I am familiar with the term “drone journalism”	46%	21%	18%	9%	6%
Q6	Citizen journalists’ contribution with multimodal data (image, video, text) makes breaking news detection easier	2%	4%	24%	47%	23%
Q7	Contribution with multimodal data (image, video, text) by citizen journalists’ that use mobile devices with geo-tagging capabilities makes breaking news detection easier	0	4%	16%	60%	20%
Q8	Citizen journalists’ contribution with multimodal data (image, video, text) makes local breaking news reporting easier	0	2%	23%	60%	15%
Q9	Citizen journalists’ contribution with multimodal data (image, video, text) leads to misconceptions	2%	15%	41%	31%	11%

In total, 60% of the respondents are not familiar with the DJ term. However, 85% of them would use an automated breaking news reporting system similar to the one presented, while 83,3% of them believe that such a system would be useful. The system is further validated by the fact that 77% believe that such a system would be useful, while 77% would use it for automated breaking news reporting. In total, 55% believe that crowdsourced data coming from citizen journalists that use mobile devices would offer better news coverage, and the same percentage (55%) believe that GPS-enabled devices would offer better news localization. This primitive user acceptance is also evaluated by the fact that 58% of the respondents do not see crowdsourced multimodal content as a threat but rather as an opportunity for better data management of breaking news.

Demographics

Out of 100 respondents, 27% are male and 73% are female. As far as the age groups to which the respondents belong, 58%, 31%, and 11% of them belong in the age group of 18–25, 26–40, and above 40, respectively. Most of the respondents’ roles (74%) are that of news consumers, while only 9% have a journalistic role (either as professional journalists or professionals in the broader mass communication field). In sum, 17% of the collected responses come from people that consider themselves as dynamic news consumers and/or distributors (via personal pages or blogs). Almost all of the respondents (99%) use the

internet at least once a day, and 78% of them use the internet for news retrieval. Figure 5 below, also presents the MEAN and Standard Deviation values for questions Q1 and Q5 – Q9, according to the results discussed above.

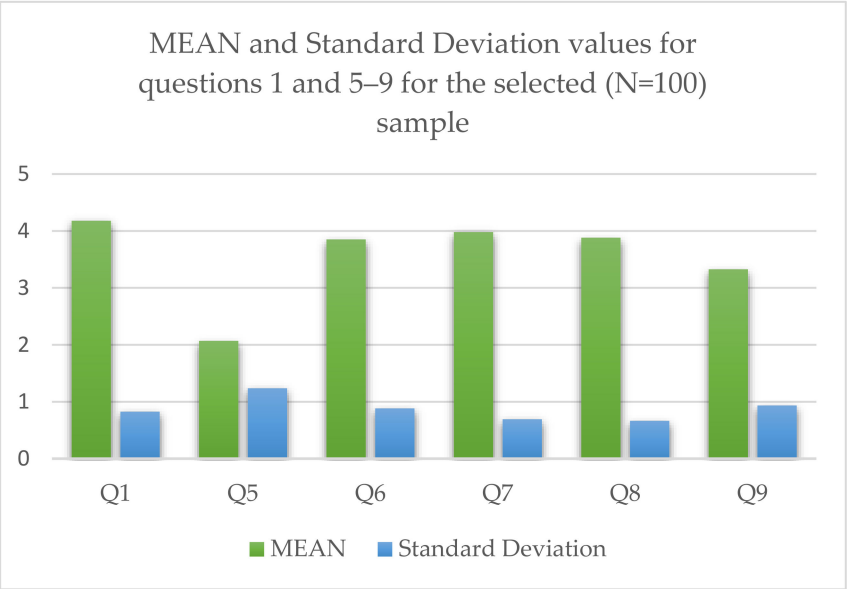


Figure 5. Graph statistics for the respondents' group (N = 100), concerning some of the important questions (namely, Q1 and Q5–Q9). It is noteworthy that although DJ is not that widespread, the participants showed a high interest in the proposed application. This can be further supported by the fact that the answers' MEAN values to the questions Q6–Q8 is near 4 ("Agree"), with S.D. values kept below 1.

In summary, the results of the conducted survey validate the research hypothesis that there is an audience willing to share (own) crowdsourced content, contributing to the easier data management of breaking news, that will provide enhanced DJ interfaces.

4.2. Qualitative System Validation Using Experts (News and Media Professionals)

While the implementation of the major components of the system was completed, the presented framework is a work in progress concerning its overall integration. Data gathering, processing, normalization, propagation to the pipeline, and storage are already in place. Moreover, similarity results are being tested end-to-end, while the algorithm to extract breaking news is under evaluation using a full load of input events. Geo-tag extraction is underway along with a flight plan and control system based on a web user interface. The aforementioned modalities and the system as a whole were validated through formative evaluation sessions, interviewing experts in the field, working in newsrooms, and manually performing most of the underlying tasks. The results are presented in Table 3 and Figure 6, in which the mean and standard deviation values of all answers are calculated. Based on these preliminary results, the usefulness of a system that automatically compares articles for the purpose of breaking news detection can be evaluated and conclusions can be extracted, utilities that most of the sample evaluators/experts would use. Lower results in the trust column are considered expected since the system is under construction and has to prove its applicability and accuracy in practice, even though there is high confidence regarding its performance in the field. It is expected that the gradual elaboration of the systems and the associated datasets would further increase the robustness and efficiency of the service in different real-life scenarios.

Table 3. Questionnaire answered, evaluating the acceptance of presented work by experts (E1–E10) (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree).

	Question	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
EQ1	Would you be interested in a system that collects all the news articles to create a data set?	4	5	4	3	3	4	5	5	4	5
EQ2	Would you be interested in a system to look up information?	2	3	2	2	3	3	3	4	4	5
EQ3	Would you be interested in a system that would allow you to compare articles with each other?	5	5	4	4	4	5	4	5	5	5
EQ4	Would you be interested in an automatic emergency detection system?	3	5	5	4	4	3	4	4	3	4
EQ5	An automatic breaking news system is useful.	3	4	4	4	4	4	4	4	3	4
EQ6	I would use an automatic emergency detection system.	3	5	5	4	4	4	4	4	4	5
EQ7	I would trust an automatic emergency detection system.	2	3	3	3	3	3	3	4	4	4

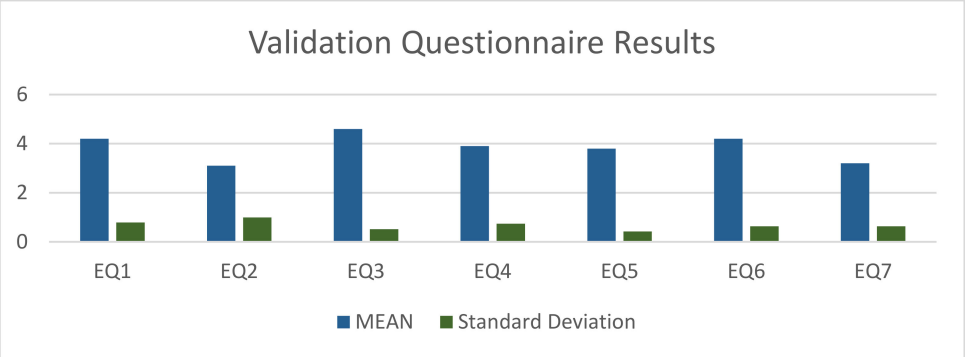


Figure 6. Mean and standard deviation of the answers given in the validation questionnaire.

4.3. Preliminary System Evaluation on Real-World Data

During the evaluation process, the need for a news database was raised. Due to the inexistence of a public and open Greek media news database, a custom one was constructed and populated, consisting of 38,759 articles, gathered in 175 days from 412 online news platforms. A custom tool for news-feed collection from real-life aggregators was developed, aiming to provide the deployed database with necessary data.

Each article was classified as "breaking" or "not breaking" by assigning the respective label to it. Classification could only be carried out by humans; thus, three experts were invited to process the article titles of two whole dates (526 articles) and place a "vote" for each one they perceived as breaking [61]. Two testing sets were created, namely *three-vote unanimous* and the *two-vote majority*. The breaking events of the former set were voted by all three experts as such, whilst the breaking events of the latter were voted by at least two experts. The dataset is publicly available on http://m3c.web.auth.gr/research/datasets/bndb/ (accessed on 9 November 2021).

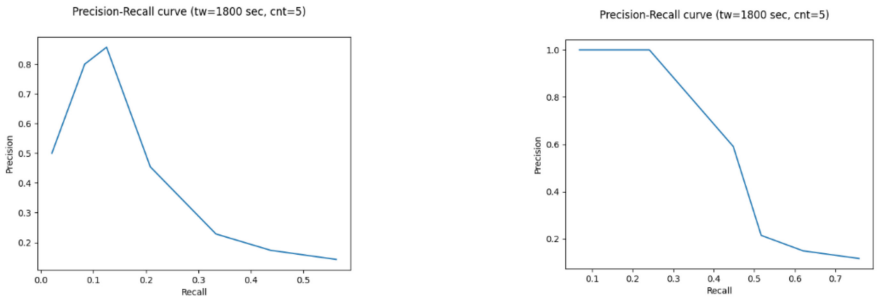
After constructing this first testing dataset and running the detection tool on it, we determined the evaluation metrics that should be used. Precision and recall are two of the most commonly used metrics in information retrieval and classification [62]. Precision is the fraction of the documents retrieved that are relevant to the user's information need:

$$\text{precision} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (2)$$

Recall is the fraction of the documents that are relevant to the query that is successfully retrieved:

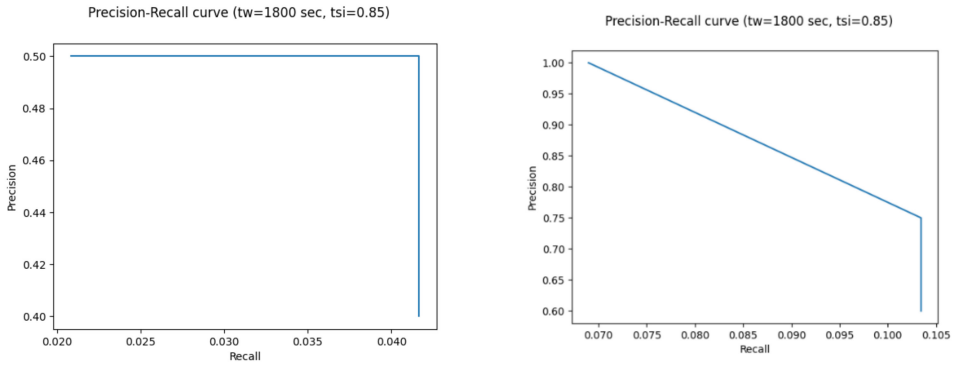
$$\text{recall} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (3)$$

The Precision–Recall curve is one of the most used metrics for the evaluation of binary classifiers. In the current use-case, this curve is affected by the involved system parameters, namely time window (*tw*) that controls observation length, similarity index threshold (*tsi*) that estimates the relation of the retrieved articles, and count threshold (*cnt*) that summates the number of the retrieved documents within the wanted *tsi* values. This was initially checked through empirical observations, with trial-and-error testing to indicate an appropriate time window of *tw* = 1800 s (i.e., half an hour). Hence, it was decided to further elaborate on the behavior of the classifier through the Precision–Recall curve concerning the change of the similarity threshold (Figure 7) and the count threshold (Figure 8). It is expected that greater threshold values will have a great impact on Precision, as the detection algorithm will be "stricter".



(**a**) (**b**)

Figure 7. Precision–Recall curve keeping the window at 1800 s and the count threshold at 5 while increasing the similarity threshold from 0.55 to 1.0 with a step of 0.5: (**a**) three-vote unanimous set; (**b**) two-vote majority set.



(**a**) (**b**)

Figure 8. Precision–Recall curve keeping the window at 1800 s and the similarity threshold at 0.85, while increasing the count threshold from 3 to 13 with a step of 1: (**a**) three-vote unanimous set; (**b**) two-vote majority set.

An attempt to evaluate the behavior of the detection system was also made by changing the time window from 600 to 6000 s while keeping the similarity threshold at 0.85 and the count threshold at 5, but the results remained the same.

Overall, based on the evaluation results of the trained models, the parameter that has the greatest impact on the classifier's behavior is the similarity threshold. Setting the similarity threshold to 0.75, a Precision value of 0.857 and a Recall value of 0.125 were achieved on the three-vote unanimous set. This leads to the conclusion that most of the retrieved articles were truly classified as breaking news, whilst the probability of a breaking article to be retrieved by the system was kept low. This prototype system performs remarkably better on the two-vote majority set, on which setting the similarity threshold to 0.75 a precision value of 1.0, and a recall value of 0.241 is achieved.

Recalling the conducted state-of-research review, related works, which deal either with hot topic detection or breaking news detection, evaluate their results with the same metrics, achieving average Precision and Recall values of 0.84 and 0.83, respectively [36,40]. It should be highlighted that the accuracy of the systems that try to solve these problems is still relatively low at the level of 80% [36,40,41]. Given that the current work faces a broader/more demanding problem, and the proposed solution fertilizes the ground for future ML/DL elaborations (the full potential of which have not been explored at this point), we can support the idea that the results achieved are mostly adequate, strongly validating the targeted proof of concept.

5. Discussion

The aforementioned preliminary results of the implemented prototype were also given to the field experts mentioned above. Based on their feedback, they are promising, but quite a few improvements are required. As the database will be growing bigger and more modules will be implemented and integrated, the results will get better. A bigger database will not only offer the opportunity of moving to deep learning methods but will also assist in the introduction of media source weights. Some media are more reliable than others, which is something that should be taken into consideration by the implemented logic. It also seems that a custom corpus should be built for training the word vector providing model, as words in the news continuously change, and a static lexicon will not be able to cover future cases.

The integration of more input sources will also improve the results, as social media will provide a greater or at least a different kind of insight than that of the websites. End users should also be able to choose specific social media accounts to monitor so as not to burden the system with a high volume of useless data.

The human factor will play a very serious role in the efficiency of this framework. Mobile journalists that will be trusted with both sending alerts to and receiving alerts from the platform will be carefully chosen. Reliability is the most important trait for both functions, which means that the selection should be made with the assistance of professionals.

6. Future Work

As soon as the breaking news detection method reaches an ideal state, the next step will be to work on the other two pillars. Flight-plan preparation demands robust and reliable geolocation inference. The next milestone will be to provide geo-location information extracted by the content of the event entities in case they are not explicitly included in them. It may be impossible to infer the exact location of interest just from textual or visual data, but humans will always be involved in the process to correct or improve it. This stage includes the trajectory design as well as logistics operations, such as drone and operator availability checking as well as requesting flight authorization if required by the law.

The last step will be the provision of semi-autonomous guiding to the drone or the human operator through waypoints. This process will demand access to three-dimensional maps and advanced image processing for obstacle avoidance.

7. Conclusions

Drones are very useful in hostile or hazardous environments where the human approach is considered dangerous and unsafe or when ground coverage is not feasible; there is also the “disposable drone scenario”, where it is expected that UAVs might not return to the land zone. Besides the aims of healthier and less dangerous working conditions for reporters, new storytelling methods can be exploited to raise public awareness on specific news stories (e.g., a conflict in a war zone, etc.) Extending the above, drone journalism could be the ideal tool for reporting physical disasters while delivering valuable civil protection informing services to the public (e.g., monitoring an earthquake or hurricane, traffic jam, etc.) Additional cases include environmental journalism, nature inspection, wildlife protection, and sports coverage.

Future drone journalism services will be able to address numerous challenges that people are already facing. For instance, environmental issues and the better management of earth resources are considered very critical for the upcoming decades. Journalists and news media are obliged to inform people and to help them realize the actual situation as well as actions that should be taken. UAV surveillance allows for long-term environmental observation, which could be quite useful in the above directions. New technologies (i.e., multispectral imaging) will allow the monitoring of invisible changes, which can be conducted regularly. While the technology of multi-spectral vision is currently expensive, it is more than certain that such imaging tools will be commonly available in the not-so-far future.

The proposed framework is in alignment with important worldwide initiatives, dedicated to establishing the ethical, educational and technological framework for this emerging field. People deserve to be properly informed and to be given further assistance on arising social, techno-economic, legal, and moral matters. Another positive impact is the implementation of smart systems and the acquisition of large-scale annotated content repositories, featuring common ground with far-reaching aims, such as the Semantic/Intelligent Web (Web 3.0/4.0) and the Internet of Things.

The benefits of such a system may be numerous, but amateur and professional users must be careful when using devices that can record video and audio. Personal data should be respected by everyone and under all circumstances. Every day, people come before cameras that are intentionally or unintentionally pointing at them. Drones have a longer range, which means that footage that includes people should be automatically censored before being publicly broadcasted. Every action should be governed by journalistic ethics.

Author Contributions: Conceptualization, M.N. and C.D.; methodology, M.N. and M.E.S.; software, M.N.; validation, M.N. and C.D.; formal analysis, M.N., C.D and A.S.; investigation, M.E.S.; resources, C.D., A.S. and A.V.; data curation, C.D., A.S. and A.V.; writing—original draft preparation, M.N. and M.E.S.; writing—review and editing, C.D., A.S. and A.V.; visualization, M.N. and M.E.S.; supervision, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data that are not subjected to institutional restrictions are available through the links provided within the manuscript.

Acknowledgments: The authors acknowledge the valuable contribution of Innews S.A. (https://www.innews.gr/, accessed on 9 November 2021) for providing us access to their repository with original indexed source data/articles that helped in proving the initial hypothesis. They also acknowledge Software Engineer Andreas Ntalakas for envisioning and setting up the system architecture design.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Drones and Journalism: How Drones Have Changed News Gathering. Available online: https://www.simulyze.com/blog/drones-and-journalism-how-drones-have-changed-news-gathering (accessed on 23 March 2017).
2. Taking Visual Journalism into the Sky with Drones. Available online: https://www.nytimes.com/2018/05/02/technology/personaltech/visual-journalism-drones.html (accessed on 2 May 2018).
3. Gynnild, A. The robot eye witness: Extending visual journalism through drone surveillance. *Digit. J.* **2014**, *2*, 334–343. [CrossRef]
4. Hirst, M. *Navigating Social Journalism: A Handbook for Media Literacy and Citizen Journalism*, 1st ed.; Routledge: Abingdon, UK, 2019.
5. How Drones Can Influence the Future of Journalism. Available online: https://medium.com/journalism-innovation/how-drones-can-influence-the-future-of-journalism-1cb89f736e86 (accessed on 17 December 2016).
6. Palino, T.; Shapira, G.; Narkhede, N. *Kakfa: The Definitive Guide*; O'Reilly: Newton, MA, USA, 2017.
7. Dörr, K.N.; Hollnbuchner, K. Ethical Challenges of Algorithmic Journalism. *Digit. J.* **2017**, *5*, 404–419. [CrossRef]
8. Ntalakas, A.; Dimoulas, C.A.; Kalliris, G.; Veglis, A. Drone journalism: Generating immersive experiences. *J. Media Crit.* **2017**, *3*, 187–199. [CrossRef]

9. Harvard, J. Post-Hype Uses of Drones in News Reporting: Revealing the Site and Presenting Scope. *Media Commun.* **2020**, *8*, 85–92. [CrossRef]
10. Virginia Tech. Mid-Atlantic Aviation Partnership. Available online: https://maap.ictas.vt.edu (accessed on 13 December 2018).
11. Valchanov, I.; Nikolova, M.; Tsankova, S.; Ossikovski, M.; Angova, S. *Mapping Digital Media Content. New Media Narrative Creation Practices*; University of National and World Economy: Sofia, Bulgaria, 2019.
12. Dörr, K.N. Mapping the field of Algorithmic Journalism. *Digit. J.* **2015**, *4*, 700–722. [CrossRef]
13. Haim, M.; Graefe, A. Automated news: Better than expected? *Digit. J.* **2017**, *5*, 1044–1059. [CrossRef]
14. Fillipidis, P.M.; Dimoulas, C.; Bratsas, C.; Veglis, A. A unified semantic sports concepts classification as a key device for multidimensional sports analysis. In Proceedings of the 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Zaragoza, Spain, 6–7 September 2018; pp. 107–112.
15. Fillipidis, P.M.; Dimoulas, C.; Bratsas, C.; Veglis, A. A multimodal semantic model for event identification on sports media content. *J. Media Crit.* **2018**, *4*, 295–306.
16. Shangyuan, W.; Edson, C.T., Jr.; Charles, T.S. Journalism Reconfigured. *J. Stud.* **2019**, *20*, 1440–1457. [CrossRef]
17. Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. [CrossRef]
18. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. 2014. Available online: https://nlp.stanford.edu/pubs/glove.pdf (accessed on 9 November 2021).
19. Katsaounidou, A.; Dimoulas, C. Integrating Content Authentication Support in Media Services. In *Encyclopedia of Information Science and Technology*, 4th ed.; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA, 2017.
20. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018.
21. Shahbazi, Z.; Byun, Y.C. Fake Media Detection Based on Natural Language Processing and Blockchain Approaches. *IEEE Access* **2021**, *9*, 128442–128453. [CrossRef]
22. Symeonidis, A.L.; Mitkas, P.A. *Agent Intelligence through Data Mining*; Springer Science & Business Media: Berlin, Germany, 2006.
23. Xiang, W.; Wang, B. A Survey of Event Extraction from Text. *IEEE Access* **2019**, *7*, 173111–173137. [CrossRef]
24. Piskorski, J.; Zavarella, V.; Atkinson, M.; Verile, M. Timelines: Entity-centric Event Extraction from Online News. In Proceedings of the Text 2 Story 20 Workshop 2020, Lisbon, Portugal, 14 April 2020; pp. 105–114.
25. Stamatiadou, M.E.; Thoidis, I.; Vryzas, N.; Vrysis, L.; Dimoulas, C. Semantic Crowdsourcing of Soundscapes Heritage: A Mojo Model for Data-Driven Storytelling. *Sustainability* **2021**, *13*, 2714. [CrossRef]
26. Chatzara, E.; Kotsakis, R.; Tsipas, N.; Vrysis, L.; Dimoulas, C. Machine-Assisted Learning in Highly-Interdisciplinary Media Fields: A Multimedia Guide on Modern Art. *Educ. Sci.* **2019**, *9*, 198. [CrossRef]
27. Drone Journalism: Newsgathering Applications of Unmanned Aerial Vehicles (UAVs) in Covering Conflict, Civil Unrest and Disaster. Available online: https://assets.documentcloud.org/documents/1034066/final-drone-journalism-during-conflict-civil.pdf (accessed on 10 October 2016).
28. Culver, K.B. From Battlefield to Newsroom: Ethical Implications of Drone Technology in Journalism. *J. Mass Media Ethics* **2014**, *29*, 52–64. [CrossRef]
29. Sidiropoulos, E.A.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C.A. Collecting and Delivering Multimedia Content during Crisis. In Proceedings of the EJTA Teacher's Conference 2018, Thessaloniki, Greece, 18–19 October 2018.
30. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C.A. A mobile cloud computing collaborative model for the support of on-site content capturing and publishing. *J. Media Crit.* **2018**, *4*, 349–364.
31. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. jReporter: A Smart Voice-Recording Mobile Application. In Proceedings of the 146th Audio Engineering Society Convention, Dublin, Ireland, 20–23 March 2019.
32. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How in Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173. [CrossRef]
33. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. Machine-assisted reporting in the era of Mobile Journalism: The MOJO-mate platform. *Strategy Dev. Rev.* **2019**, *9*, 22–43. [CrossRef]
34. Petráček, P.; Krátký, V.; Saska, M. Dronument: System for Reliable Deployment of Micro Aerial Vehicles in Dark Areas of Large Historical Monuments. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2078–2085. [CrossRef]
35. Krátký, V.; Alcántara, A.; Capitán, J.; Štěpán, P.; Saska, M.; Ollero, A. Autonomous Aerial Filming With Distributed Lighting by a Team of Unmanned Aerial Vehicles. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7580–7587. [CrossRef]
36. Bok, K.; Noh, Y.; Lim, J.; Yoo, J. Hot topic prediction considering influence and expertise in social media. *Electron. Commer Res.* **2019**, *21*, 671–687. [CrossRef]
37. Liu, Z.; Hu, G.; Zhou, T.; Wang, L. TDT_CC: A Hot Topic Detection and Tracking Algorithm Based on Chain of Causes. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*; Springer: Cham, Switzerland, 2018; Volume 109, pp. 27–34. [CrossRef]
38. Hoang, T.; Nguyen, T.; Nejdl, W. Efficient Tracking of Breaking News in Twitter. In Proceedings of the 10th ACM Conference on Web Science (WebSci'19), New York, NY, USA, 26 June 2019; pp. 135–136. [CrossRef]
39. Shukla, A.; Aggarwal, D.; Keskar, R. A Methodology to Detect and Track Breaking News on Twitter. In Proceedings of the Ninth Annual ACM India Conference, Gandhinagar, India, 21–23 October 2016; pp. 133–136. [CrossRef]

40. Jishan, S.; Rahman, H. Breaking news detection from the web documents through text mining and seasonality. *Int. J. Knowl. Web Intell.* **2016**, *5*, 190–207. [CrossRef]
41. Zhu, Z.; Liang, J.; Li, D.; Yu, H.; Liu, G. Hot Topic Detection Based on a Refined TF-IDF Algorithm. *IEEE Access* **2019**, *7*, 26996–27007. [CrossRef]
42. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv* **2021**, arXiv:2004.03705. [CrossRef]
43. Xu, G.; Meng, Y.; Chen, Z.; Qiu, X.; Wang, C.; Yao, H. Research on Topic Detection and Tracking for Online News Texts. *IEEE Access* **2019**, *7*, 58407–58418. [CrossRef]
44. Web Scrapping Using Python and Beautiful Soup. Available online: https://towardsdatascience.com/web-scraping-5649074f3ead (accessed on 20 March 2020).
45. Avraam, E.; Veglis, A.; Dimoulas, C. Publishing Patterns in Greek Media Websites. *Soc. Sci.* **2021**, *10*, 59. [CrossRef]
46. Dean, A.; Crettaz, V. *Event Streams in Action*, 1st ed.; Manning: Shelter Island, NY, USA, 2019.
47. Psaltis, A. *Streaming Data*, 1st ed.; Manning: Shelter Island, New York, NY, USA, 2017.
48. Papadopoulos, S.; Datta, K.; Madden, S.; Mattson, T. The TileDB array data storage manager. *Proc. VLDB Endow.* **2016**, *10*, 349–360. [CrossRef]
49. TileDB. Available online: https://docs.tiledb.com/main/ (accessed on 23 January 2021).
50. Guo, W.; Zeng, Q.; Duan, H.; Ni, W.; Liu, C. Process-extraction-based text similarity measure for emergency response plans. *Expert Syst. Appl.* **2021**, *183*, 115301. [CrossRef]
51. Yang, S.; Huang, G.; Ofoghi, B.; Yearwood, J. Short text similarity measurement using context-aware weighted biterms. *Concurr. Comput. Pract. Exp.* **2020**, e5765. [CrossRef]
52. Shahmirzadi, O.; Lugowski, A.; Younge, K. Text Similarity in Vector Space Models: A Comparative Study. In Proceedings of the 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 17 February 2020; pp. 659–666. [CrossRef]
53. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
54. Azunre, P. *Transfer Learning*; Manning: Shelter Island, NY, USA, 2021.
55. Bodrunova, S.S.; Orekhov, A.V.; Blekanov, I.S.; Lyudkevich, N.S.; Tarasov, N.A. Topic Detection Based on Sentence Embeddings and Agglomerative Clustering with Markov Moment. *Future Internet* **2020**, *12*, 144. [CrossRef]
56. Middleton, S.E.; Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, Y. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Trans. Inf. Syst.* **2018**, *36*, 40. [CrossRef]
57. Dong, W.; Wang, Z.; Charikar, M.; Li, K. High-confidence near-duplicate image detection. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 5–8 June 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 1–8.
58. Li, X.; Larson, M.; Hanjalic, A. Geo-distinctive visual element matching for location estimation of images. *IEEE Trans. Multimed.* **2017**, *20*, 1179–1194. [CrossRef]
59. Li, Z.; Shang, W.; Yan, M. News Text Classification Model Based on Topic model. In Proceedings of the IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016; p. 16263408. [CrossRef]
60. Patel, S.; Suthar, S.; Patel, S.; Patel, N.; Patel, A. Topic Detection and Tracking in News Articles. In Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, 25–26 March 2017. [CrossRef]
61. Dimoulas, C.; Papanikolaou, G.; Petridis, V. Pattern classification and audiovisual content management techniques using hybrid expert systems: A video-assisted bioacoustics application in Abdominal Sounds pattern analysis. *Expert Syst. Appl.* **2011**, *38*, 13082–13093. [CrossRef]
62. Rinaldi, A.M.; Russo, C.; Tommasino, C. A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features. *Future Internet* **2020**, *12*, 183. [CrossRef]