International Journal of Geo-Information

MDPI

*Article*

# Using HyperLogLog to Prevent Data Retention in Social Media Streaming Data Analytics

**Marc Löchner * and Dirk Burghardt**

Institute of Cartography, TU Dresden, Helmholtzstr. 10, 01062 Dresden, Germany
* Correspondence: marc.loechner@tu-dresden.de

**Abstract:** Social media data are widely used to gain insights about social incidents, whether on a local or global scale. Within the process of analyzing and evaluating the data, it is common practice to download and store it locally. Considerations about privacy protection of social media users are often neglected thereby. However, protecting privacy when dealing with personal data is demanded by laws and ethics. In this paper, we introduce a method to store social media data using the cardinality estimator HyperLogLog. Based on an exemplary disaster management scenario, we show that social media data can be analyzed by counting occurrences of posts, without becoming in possession of the actual raw data. For social media data analyses like these, that are based on counting occurrences, cardinality estimation suffices the task. Thus, the risk of abuse, loss, or public exposure of the data can be mitigated and privacy of social media users can be preserved. The ability to do unions and intersections on multiple datasets further encourages the use of this technology. We provide a proof-of-concept implementation for our introduced method, using data provided by the Twitter API.

**Keywords:** social media; privacy protection; data retention; disaster management; geocode systems; privacy-aware data storage; cardinality estimation; hyperloglog algorithm; Twitter



**Citation:** Löchner, M.; Burghardt, D. Using HyperLogLog to Prevent Data Retention in Social Media Streaming Data Analytics. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 60. https://doi.org/10.3390/ijgi12020060

Academic Editors: Wolfgang Kainz, Maria Antonia Brovelli, Songnian Li and Ivana Ivánová

Received: 22 December 2022
Revised: 30 January 2023
Accepted: 6 February 2023
Published: 9 February 2023



## 1. Introduction

Social media services are ubiquitous and often faster than conventional media in terms of information distribution. This advantage is made use of in numerous situations, for example in disaster management [1]. Modern disaster management concepts, such as *Virtual Operation Support Teams* (VOSTs), even depend on this kind of information source. For them to be able to accomplish their tasks, social media data are being gathered and stored, before being analyzed and evaluated.

Analyzing and evaluating social media data by counting appearances of posts that contain certain information is a common practice. For example, the topic of their payload content, but also their attached location data, date and time, or even information about the user are valuable criteria to search for. The result then is a list of posts matching these search criteria that can be used as a basis for diagrams or overview maps highlighting trends or hot spot areas.

A common approach to gather the data is to query public interfaces provided by social media services for either a real-time data stream or a historical list of posts, and store the resulting data in local databases to be processed by analytics applications in order to investigate their characteristics. This approach is called *exploratory data analysis*, but can also be interpreted as *data retention* (see Section 2.2), a practice which is problematic in numerous ways. Recent incidents of contact tracing data misuse [2] stand exemplarily for side-effects of gathering large sets of personal data. Once a set of data has been gathered for whatever reason, it is subject to being misused by third parties. Especially in disaster management, personal data must be taken good care of. People being affected by disasters are vulnerable. They are potentially dependent on receiving or sharing reliable information or seeking help through social media services. It is even possible that using social media

services is their only opportunity. This makes retention of this kind of data problematic in terms of their original creators' informational self-determination: once data are collected and stored by a third party, their original creator can neither update nor delete it and, thus, has lost control over it. Moreover, having access to such datasets enables to extract a list of users who have obviously been at a certain place. This can be crucial in relevant situations with special privacy interest for users, for example in refugee movements, demonstrations, or riots.

This raises the question of the necessity to store large amounts of personal data from the social media services. A progressive and privacy-aware approach to answer this question is to not store *collateral data*, meaning that kind of data that are not necessary for the accomplishment of the task that the data are being gathered for. In our research, we aim to provide methods and technology to process social media data following this approach. We propose to actively prevent the gathering of collateral data and, thus, better protect privacy of social media data creators.

To achieve this, we introduce a method to store social media data in a structure, that is build on top of a data storage algorithm called *HyperLogLog* (HLL). We show how to use the technology on real-time streaming social media data with a usage scenario in disaster management. This is especially significant, because the usefulness of such data is very ephemeral. Once a disaster is overcome, most of the data have lost their relevance. This increases the urgency to pay attention to what data need to be gathered in the first place.

Our concept is developed with generic application on any social media services or networks in mind. For the presentation in this paper, we focus on data taken from the well-known social media service Twitter. We reference Twitter posts as our example social media data, which we accessed through the Twitter API with academic research access level [3]. Readers should be aware tough, that Twitter is only used exemplarily as a data source and should not be considered an obligatory foundation for this work. The herein presented concepts apply to data sourced from any social media platform, including decentralized networks such as the fediverse [4].

In Section 2, we explain the fundamentals on disaster management in general, the concept of VOSTs, why data retention is a serious threat and the fundamental functionality of HLL. Following up in Section 3, we describe our proposed concept to store social media data, without accidentally storing collateral data, by utilizing HLL to mitigate the data retention threat. We also describe a potential scenario in disaster management, wherein the concept is applied. In Section 4, we will give a short insight in the proof-of-concept implementation. Afterwards, we discuss the pros and cons of our proposed method in Section 5, concluding with an outlook of further research.

## 2. Fundamentals

### *2.1. Disaster Management*

The ubiquity of social media services has made them an established interactive communication platform for people worldwide. Today, they are a crucial data source in disaster management. People actively use them to obtain information about incidents, extent of damage, possible further dangers, or to keep in contact with their relatives and offer help [5]. The public traces of their social media usage is an enormous pool of valuable information for situation assessment and rescue operations. Strategies to utilize this information are subject to a wide range of research. Public organizations, such as Digital Humanitarian Networks, emerge as new ways of digital organization structures, which enable new forms of engagement [6]. A guided and updated bibliography on human-centered research in crisis informatics [7] further outlines the scope of the topic.

Flood incidents mark special scenarios, in which social media data are crucial to handle the situation. Khan et al. [8] show exemplarily that in certain situations, social media data are even more reliable than real-time flow gauges data. While they do machine learning on imagery and, thus, focus on assessing specific situations, Barker and Macleod [9] prototyped flood-related Twitter data mining on a national scale. Even larger, de Bruijn et al. [10] show

a global database of historic and real-time flood events based on social media. None of these projects consider privacy issues within their research.

Fathi et al. [11] describe a digital help concept around *Virtual Operations Support Teams* (VOSTs). These teams consist of groups of volunteers, who come together in times of crises, disasters, or other potentially dangerous situations, such as demonstrations or events. Their task is to constantly monitor social media services for any information that is relevant to discover, assess, or mitigate dangerous situations and report to decision makers of public authorities. However, protecting the privacy of social media users is not considered a primary focus here either. In this paper, we will use this example of the subsequent processing of social media data as an example scenario.

### 2.2. Data Retention

The work of VOSTs stands exemplarily for data analyses that utilize analytics software, which provides contextual overviews on previously gathered data stored in local databases. User interfaces take input to be crawled for in the stored data and return statistics of, e.g., occurrences in their topical, spatial, temporal, or social facets. This can be the topic of its payload text, the attached location data, date and time of its publication, or information about the creators and who they follow. The process follows the concept of *exploratory data analysis*, first described by Tukey [12], which can be summarized by "store everything possible and look for interesting information in it afterwards". Depending on the scenario, only parts of the stored information may be relevant for the originally intended analysis, in this scenario "look for potential disaster indications". For example, knowing the user names of the posts is not necessary for this analysis. Still, the entirety of every post has been captured from the social media service. That means that if a post is being altered or deleted on site of the social media service, it still resides in the original form at the place where it has been downloaded to. Technically, that practice meets the requirements to be termed *data retention*. Our understanding of the term is preserving data for an indefinite time period with no specific purpose for any individual data item, but with the perspective to make use of the information partially or in its entirety at a later point in time.

The term *data retention* is being discussed in the public mostly in conjunction with surveillance of public telecommunication usage [13]. The European Digital Rights public interest group states that "data retention practices interfere with the right to privacy at two levels: at the level of retention of data, and at the level of subsequent access to that data by law enforcement" [14]. In this paper, we make use of that term's alarming connotation and introduce it in a broader and more technical environment. Doing so, we want to emphasize the explosive practice of recklessly dealing with personal data. According to the above definition, the use of the term is valid for any case of storing and retending personal data in stocks. Wright et al. [15] use it even to describe any storage of data underlying scientific studies. Being in possession of personal data requires great responsibility in terms of data security, as it opens up risks of possible abuse, theft, or accidental public exposure [16]. Guillou and Portner [17] break it down to the simple rule "the more data you have, the more data you can lose".

Players beyond governmental agencies and law enforcement, such as journalists, researchers, or non-profit organizations, face even more challenges when dealing with social media data. Stieglitz et al. [18] point out that the volume of data was most often cited as a challenge by researchers. Wang and Ye [19] coin the term *mining* when summarizing common techniques for social media analytics in natural disaster management.

The social impact of misusing large sets of data is well-known. Blanchette and Johnson [20] explain how data retention threatens the social concept of *forgetfulness*. Recent incidents of contact tracing data misuse of German Police [2] and the Cambridge Analytica scandal [21] are two examples of how retended data can be misappropriated. Incidents such as these make users of social media services start to realize that all of their data are not just used for their initial purpose. They learn that there are entities out there who have the power to gain access to gathered datasets and the impertinence to abuse them. The *Chilling*

*effect* is used to describe the consequences of peoples slowly increasing self-discipline and restriction of their communication behavior due to becoming aware of digital surveillance and panopticism [22], as found by Büchi et al. [23]. It is of concern that people tend to retreat from public social media services in favor of closed, "antisocial" messaging groups [24], Wilson [25].

However, social media data can not only be used to gain knowledge and, thus, power over society. Social media data analytics in disaster management is an example for positive reuse of social media data, but also an example for the direct dependence on it. The work of humanitarian organizations relies on publicly available data, that is authentic and relevant. In particular, VOSTs depend on public availability of social media data [26]. Therefore, the gradual retreat of users from social media services must be prevented.

### 2.3. Geocode Systems

The concept presented in this paper includes utility of the geocode identification concept *geohash* [27]. Similar to a quad tree [28], a geohash is a spatial data structure used to represent a certain area on the globe in an alphanumeric representation. The structure is based on a hierarchical discrete grid of four areas, that alternately follow the *Z-order-curve* function [29]. The size of the area depends on the precision specified by the length of the geohash.

There is a large range of geocode systems available. While many of them are tied to a certain scope, e.g., postal codes, there is a list of general scope systems. Among those, some of them are restricted by patents or tied to administrative divisions, defining, e.g., only country codes, such as `DE` or `UK`, which lacks enough flexibility to define area sizes. Within the remaining general grid-based geocode systems are geohashes, Google's "Plus codes" [30] and Yahoo's now deprecated "Where on earth ID" [31]. Because their representation is a simple string, their creation unambiguous and their implementation independent of a private company, we decided to use geohashes as geocode system for the representation of spatial information in our concept.

### 2.4. Privacy-Aware Data Storage

A common method to store data without leaking information about individual items is *Differential Privacy* (DP) [32]. In its foundation, it adds randomness to a dataset. Random data are indistinguishable from the real data and, therefore, cloaks the real data within the set. The method though requires statistical knowledge about the scope of the real data, in order to define the distribution of randomness [33]. With streaming data it is impossible to make assumptions about its scope, since we can not look into the future. Another drawback of DP is that it increases the size of the dataset, which would impair the processing performance, as social media datasets are usually large by themselves.

Many other suggestions to the problem of privacy-aware data storage exist. Most of them target the problem of preventing the storage provider from accessing the data. Suggested approaches include encryption and fragmentation techniques [34–36]. The goal of these approaches is to still have access to the original data if being in possession of decryption keys. This carries the risk of exploitation by unauthorized personnel. Exemplary scenarios range from accidentally pushed private keys to public repositories [37] to an issue of a National Security Letter [38].

Our contribution to the issue of privacy-aware social media data storage focuses on storing data without the subsequent ability to access individual items. It is based on an algorithm called *HyperLogLog* (HLL), which is a cardinality estimator first presented by Flajolet et al. [39]. Its fundamental strength is the ability to *estimate* the distinct count of a multiset (its cardinality), and store it in a data structure, that does not allow the extraction of individual elements. This is performed by storing only hashes of data items instead of the original raw data and identifying them by counting leading zeros of the binary representation of their hashes. The algorithm is able to predict how many distinct items have been added to the HLL set, based on the maximum number of leading zeroes

observed. This makes processing data using HLL very efficient in terms of processing time and storage space.

Furthermore, it needs external knowledge to be able to identify the existence of single items in the dataset [40]. For example, if the cardinality of a set does not change after adding an item to it, it is obvious that it has already been inside the set. However, this requires the knowledge of the item in beforehand. It is not possible to search for prior unknown information in an HLL set, such as, for example, the user names of all the posts that have been gathered.

The original HLL algorithm has been shown to have security issues and other shortcomings that can lead to large estimation errors or allow attacks on its result. However, the authors also presented mitigation methods for the addressed issues and further improvements to the algorithm [41–44].

In a case study carried out as a focus group discussion with VOST members [45], we identified HLL as a valid and privacy-aware alternative to storing raw data in the presented scenario. Within that publication, we also compared computed cardinalities of HLL sets with the actual count values in the original data. Dunkel et al. [46] agree that the overall accuracy is affected by several parameters, including the size of the data: the larger the set, the more accurate the estimation. Desfontaines et al. [40] show that the estimation accuracy directly relates to the overall privacy-preserving effect of HLL.

## 3. Concept

Based on pre-described fundamentals (see Section 2), our previous work on HLL [46,47] and findings resulting from our case study [45], we present a concept to store and process data from a real-time stream of social media posts in a way that protects from the side-effects of data retention.

In addition to its unknown extent, the particular characteristic of real-time streaming data is its ephemerality. Data come in and can be processed, but the next moment they are gone and cannot be reconsidered, unless they are stored locally. However, according to our research goals, the storage of collateral data should be prevented.

In this section, we first outline the term *collateral data*, as well as the purpose of geohash locations in the concept. We then introduce the utility of the HLL technology to only store necessary data. Finally, we classify the concept using an operational scenario.

### *3.1. Collateral Data*

When utilizing some analytics software to monitor the occurrence of pre-defined terms, conventional analytics tools would store the social media data in a database, to be able to run further analytics and visualizations over it. Typically, relational or non-relational database management systems are used ([18] p. 163). This procedure brings the side-effect of storing *collateral data*, which are data that are not required to fulfil the task of computing the post cardinality. An example of a data item for the occurrence of a well-known social media post matching a search criteria is shown in Figure 1b. In addition to the location of the post, it also includes information about the author, the date and time, as well as the entire content of the post. These data can be used for example to identify the original creator of the post, and, thus, for other purposes than originally intended, which we identify as data retention (see Section 2.2).

To determine the cardinality of posts containing the term `flood` in a certain area, most of the data shown in Figure 1b are not required. We know that the present post matches our pre-defined term, because we had created a search rule for it (see Section 4) and, therefore, all posts we receive will match the search term. The actual content of each post is not necessary to store, consequentially. Furthermore, we neither require the time of creation nor the author for each individual post to state the cardinality of all the posts.

In order to state the cardinality of posts, we do need to assign the posts some ID as the unique identifier of each post. In our proof-of-concept implementation (see Section 4), such an ID, is already provided by the Twitter API (red color in Figure 1). We also need the

location data of each post (green color in Figure 1) in order to determine the location of the potential flood incident. For any location-based social media data analysis, e.g., a VOST to be able to localize potential disaster situations, the need for some sort of geo-referencing data is essential. Our concept introduces a method to store these data in a way that it is only useful to determine the post cardinality, the number of posts with occurrences of said pre-defined terms.

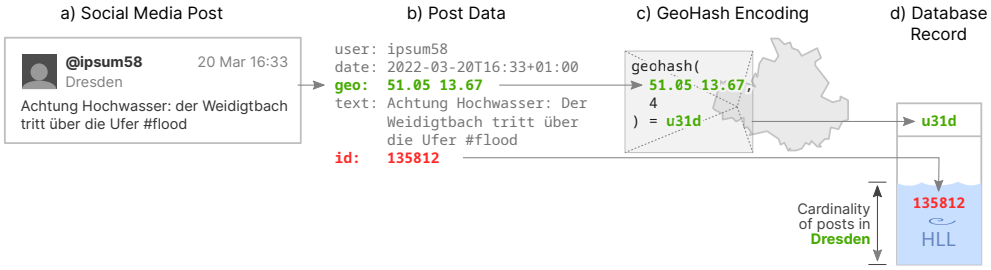


**Figure 1.** Social media data processing graph. (**a**) Example post. (**b**) The post's social, temporal, spatial (green), and topical data, and its hidden unique ID (red). (**c**) Encode the corresponding geohash from the geo-coordinates. The result represents the area plotted by the rectangle over the outlines of Dresden. (**d**) Store the post ID in the HLL set of the database record matching the geohash.

### *3.2. Geohash Locations*

The concept is about determining the cardinality of posts for a certain area. To sort individual posts into areas, the geo-location of a post is generalized by converting the original location data to a *geohash* (see Section 2.3). Posts come with geo-location information of multiple quality levels (see Section 4), each still representing a point value (latitude, longitude). The concept requires the geodata to be converted to an area, so that multiple posts can be associated with it. If each post had its own location point value, there would be an individual database entry for every post (see Section 3.3), unless multiple posts had been sent from the exact same place. Therefore, a suitable precision value must be defined for the geocode along with the search term in our concept. If the precision is too low and, thus, the area too large, there will be too many posts and potential incidents are more difficult to locate. If the precision is too high and, thus, the area too small, there will not be enough posts in an area to be able to determine anomalies in their occurrence. In Figure 1c, the precision value is 4, so the resulting geohash has a length of four characters.

### *3.3. HyperLogLog Storage*

We declare the spatial information (represented by a geohash) as the key characteristic of a social media post. It is stored into the database in clear text, serving as the index of the database record (see Figure 1d). The identifier (ID) of the post is in turn stored in a HyperLogLog (HLL) set in relation to its geohash. Post IDs that arrive later in the stream and match the same geohash will be added to this HLL set, which increases its cardinality by one for each new post. The resulting HLL data structure represents all posts matching a certain term from a certain area, from which it is impossible to derive the post IDs back from it. Figure 1d depicts this procedure, showing the post ID "plunge" into the HLL set (represented by a basin or sink).

Utilizing HLL, we do *not* store the post IDs itself, but only calculate hashes from them and store them in an array of counters that represents the set of post IDs. Figure 2 shows an exemplary database table structure with geohashes and post IDs. The geohash values each represent an area, and the corresponding HLL sets represent the IDs of posts that occurred in that area. Having a database with geohashes and their corresponding HLL set, as shown in Figure 2, it is possible to compute the cardinality of the HLL set and, thus, determine the number of posts in each area.

| ***flood*** | |
|---|---|
| geohash | id |
| w41s | \x128b7fdf939b45ec2ef0ca |
| 6yws | \x128b7fbfd17eca803517d2 |
| c29s | \x128b7fe00ef312fcf023c9 |
| 75cs | \x128b7fcc47a6c00361c5e7 |

**Figure 2.** Exemplary structure of a database table that stores all data referring to the pre-defined term `flood`. It shows four records, each stands for one area represented by the geohash, and the corresponding HLL set containing the post IDs.

The result of a set's cardinality computation could as well be achieved by just incrementing an integer per seen post ID and storing the sum instead of an HLL set. The significance about using the HLL algorithm instead, is that it provides the opportunity to do *set operations*, such as unions and intersections, on the HLL sets. Both operations allow quantitative evaluations of relationships between HLL sets and can useful for combinations of multiple individual sets. Additionally, sets of multiple terms can be combined. The result can, e.g., support VOSTs in specifying a disaster scenario (see Section 3.4). An intersection of `fire` and `forest` sets could lead more precisely to disaster incidents than both terms on their own. It still makes sense to monitor the terms individually in the first place, because, for example, a combination of `fire` and `accident` can lead and different incidents than a combination of `forest` and `accident`.

Furthermore, different terms could have the same meaning, for example `flood`, `high tide`, `wave`, and `tsunami` could all refer to the same situation. So a union of HLL sets on posts over these terms can increase the accuracy of a disaster detection (see Section 3.4). Likewise, terms in different languages could also be monitored in combination. This may, e.g., enable VOSTs to monitor larger, multiple languages involving areas, such as border triangles, or including smaller countries, such as Benelux or the Baltics. Figure 3 emphasizes that the combination of different terms can lead to more accurate visualizations and, therefore, more rational assessments of the situation.


**Figure 3.** Visualizing the cardinality of posts containing (**a**) the term `omicron` spelled with a C, (**b**) `omikron` spelled with a K, and (**c**) the union of the sets, per area defined by geohash precision 4. The union of the two datasets helps to understand that the early 2022 variant of COVID-19 is a trending topic in more areas than what a consideration of each individual set suggests. Data from Twitter, January through March 2022. Classification: Head/Tail Breaks [48].

### 3.4. Scenario

In this subsection, we present our concept in an exemplary disaster management scenario, in which a VOST passively monitors a real-time social media data stream (see Section 2.1). A software system does the monitoring in a day-to-day routine and notifies about peaks or anomalies in order to detect potential disasters. The VOST would define a set of terms, which indicate potential disasters and, therefore, should be monitored for occurrence within posts. There is an expected noise floor, namely an average occurrence of these terms during normal times. For example, every hour there are between 50 and 500 posts in a certain area containing the term `flood`. The number of posts will normally increase by a value in this range per hour. In our concept, we call the number of posts their *cardinality*.

During a disaster, the occurrence of posts, including certain terms, and thus their cardinality may increase by a recognizably higher amount than the average, for example by 1238. For our concept, we propose the utility of a monitoring tool that checks the cardinality in a defined time range, for example every hour. Visualizing the *gain rate* of the post cardinality will result in a peak in the graph (see Figure 4). If the peak exceeds a pre-defined threshold, it will trigger an alert. It may indicate a potential disaster situation and a need for attention by the VOST.

Figure 4 shows an exemplary graph of the post cardinality gain over time and its corresponding gain rate. It includes a pre-defined threshold of posts per time range set to 500. At position 4, a large cardinality gain by 1238 posts occurs, which exceeds the threshold. It triggers an alert for the VOST with the corresponding monitored term and the threshold, for example "term `flood` has exceeded 500 posts per hour".

| Hours | Cardinality | Gain Rate |
|---|---|---|
| 0 | 58 | 58 |
| 1 | 109 | 51 |
| 2 | 184 | 75 |
| 3 | 345 | 161 |
| 4 | 1583 | 1238 |
| 5 | 2036 | 453 |
| 6 | 2356 | 320 |
| 7 | 2495 | 139 |
| 8 | 2528 | 33 |
| 9 | 2607 | 79 |

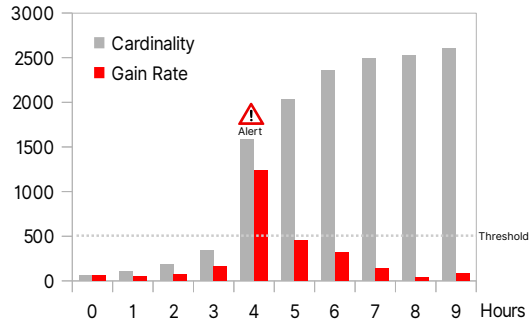


**Figure 4.** Post cardinality gain. The table on the left shows the cardinality gain per hour and its rate accordingly. The chart on the right shows the chart visualization of the table. It outlines the threshold through the dotted line at 500 posts per hour. The attention sign highlights the exceeding cardinality gain rate and, thus, marks the point of alert.

Once the VOST is alerted for a peak in the cardinality gain for some pre-defined terms, they may start their investigation on that matter. This usually includes browsing social media services through their search functions in order to receive live occurrences of posts relevant to the context [11]. This stage marks the end of our concept scenario.

## 4. Implementation

We created a proof-of-concept implementation for this concept called *VGIsink* [49]. It is designed as an HTTP-based RESTful API [50] to ensure standard-compliant access for any client application. While the proposed concept is dedicated to work generically with data from any social media platform, the VGIsink implementation is restricted to compatibility with Twitter.

Setting up a working example requires to gain access to the Twitter API in order to curate a custom real-time stream of posts. The *filtered stream* feature provided through the Twitter API [51] allows to define *rules* with terms to be monitored to curate such a custom

stream. Adding the `has:geo` operator to each term ensures that the stream is limited to posts that contain geo-information. Furthermore, the geohash precision value must be defined for each rule, in order to set the size of the areas in which post cardinalities should be computed. This is out of scope of the Twitter API and, thus, we made it part of VGIsink.

Within our implementation, we adopted the term *rule* from the Twitter API to define an individual target to monitor. A rule is defined by a *term* and a *precision value*. For each rule, a table is created in the application database, following the structure described in Figure 2. Every arriving post in the real-time stream is defined by its geo-information. It can be either a specific coordinate with a latitude and a longitude value or a bounding box describing a more generic place. In case of a specific coordinate, it will be transferred to the corresponding geohash according to the defined precision value. A bounding box will be transferred into a list of all geohashes, whose center is inside the bounding box. We assume relevance of a post for the entire place, so the cardinality will increment for each geohash in that list. The geohash will be stored in clear text as the table record's primary key. The ID of the post will be added to an HLL set in the table record corresponding to the geohash. The VGIsink implementation utilizes the PostgreSQL HLL implementation [52].

Reading the resulting data means querying a certain VGIsink rule. A query to such a rule returns a JSON-formed list of areas and their corresponding cardinality. Each area is converted from the geohash to a Postgis geometry [53] and then returned as a standard GeoJSON [54] compliant coordinate, as shown in Listing 1. The cardinality is calculated using the according PostgreSQL HLL function.

A list of areas with their corresponding cardinality can be visualized for example in a mapping application. Figure 3 shows an example implementation using Leaflet [55]. It features a number of rectangular areas, each representing a geohash. The color of a rectangle represents its cardinality, where lighter means lower and darker means higher values.

**Listing 1.** Example of a list of GeoJSON objects with the corresponding cardinality.

[ "type": "Polygon", "cardinality": 108, "coordinates": [ [ [ 5.625, 49.21875 ], [ 5.625, 50.625 ], [ 7.03125, 50.625 ], [ 7.03125, 49.21875 ], [ 5.625, 49.21875 ] ] ] ]

## 5. Discussion

The rationale behind the concept presented in this paper is to provide a concept to utilize the HLL technology in order to prevent unnecessary data retention. In the following we evaluate the concept, discuss its scope and go into a number of potential alternative approaches.

The conventional way to store social media data prior to analytic processes is to just store the raw data without any further preparation. This, of course, leads to data retention as the major point of criticism (see Section 2.2). With DP we already named a potential alternative strategy to store data in a privacy-aware way in Section 2.4, though we have also declared it unsuitable in the scenario of storing social media data for its characteristics of enlarging the already large amount of data even further. In addition, defining the distribution of randomness is not possible, if the scope of the data is unknown, which is true for a stream of data whose end is indefinite.

A trivial alternative would be to only store the number of posts matching a certain rule as an integer in the database. This would not only reduce the data footprint immensely and prevent the possibility to make statements about individual items within the set. However, it only allows basic arithmetic functions and takes away the opportunity to do set operations, such as unions and intersections, over multiple datasets. We do not claim that using HLL to solve the problem is the only alternative. The low storage footprint and the fast processing speed make it a very suitable method to process social media data with their usual characteristics of being really extensive.

While we introduced the concept using only minimal examples in Sections 3 and 4, it is actually intended to operate on a much larger number of terms and rules, respectively. It could be hundreds or even thousands of them, the number of rules is potentially unlimited.

It is possible to consider only posts for one HLL set, that contain two or more specific terms, or if a combination of events from different facets occur, e.g., a term occurs during a pre-defined time range. Search terms are also not limited to nouns as introduced, but they can include verbs, adjectives, or any other kind of words. As described in Section 3.3, the combination of terms may sharpen their semantics and enables a more precise dataset. Preselecting relevant terms could be automated by a suitable topic modeling technique [56], to find terms with similar meanings automatically. Nevertheless, choosing the right terms to redeem good results needs the experience of professionals, which members of VOSTs are expected to bring along. If new topics arise on social media, new terms appear along, e.g., `covid`, new rules must be created for each.

A considerable extension to this concept might feature automatic adjustments of the geohash precision value, e.g., once a certain cardinality is exceeded. A higher precision value would split the area into smaller pieces and, therefore, reduce the size of each area. Resulting HLL sets can then be unified or intersected with other sets on their own. This argumentation is only theoretical, and no experiments have been performed on this approach. The concept relies on the experience of VOSTs to determine a sane value.

The most important limitation of the concept is the lack of ability to perform exploratory or otherwise selective qualitative analysis. It is not intended by design to be able to investigate for clues within a dataset, that have not been planned to discover. For further investigation on individual incidents that this concept can detect, VOSTs derive to other tools or applications provided by social media platforms themselves anyway [45].

With respect to common facets [57], this concept only considers spatial information in terms of collected data. In combination with other data spatial information is deemed privacy-relevant [58]. However, as shown in Figure 2, spatial information in the form of the geohash is the only data stored in clear text. The social facet, information about the user, is not stored as per definition of the concept. Temporal information can only be retrieved, if cardinalities are queried periodically and stored, e.g., in a time series database, such as Prometheus [59] or InfluxDB [60].

The ID of a post stored in an HLL set along with the corresponding geohash represents the entire social media data item. It can not be retrieved from the HLL set per definition of the HLL algorithm. However, if the ID of a post is known, it is possible to evaluate, whether it is included in an HLL set. An attacker only needs to calculate the cardinality of the set, then add the post ID to the set and then calculate the cardinality again. If it has not changed, then the post ID has been in the set before. This shows that HLL itself cannot preserve privacy, if the attacker has further information [40]. It does not impair the concept though, because the described situation is not considered an attack vector. Attempts to discover single items in an HLL set of social media post IDs is regarded unnecessary effort. Since the data are publicly available from the social media services, attackers can also obtain them from there directly. An exception to this may be a case in which a social media post has been deleted online, while its post ID is still known to the attacker. However, this will only prove that the post has existed. This is trivial, if the attacker is already in possession of the ID. No other content from the post can be recovered from the HLL set. Adversarial perturbations of the input stream to alter the cardinality estimation of an HLL set through the exploitation of security flaws (see Section 2.4) are also of purely theoretical use. This concept aims to prevent data retention and attacks to the entirety of a dataset. For example, it prevents revealing all the user names, that have posted in a certain area. This can be crucial in relevant situations of disaster management. Exemplary situations with special privacy interest for users include refugee movements, demonstrations, and riots, among others.

Edge cases involve situations, in which, for example, there is only one post within a certain area. It is obvious to unveil its identity, if the cardinality of a geohash is 1 and the attacker can look up the social media service for posts within that area, assuming it has not been deleted by the time. This can be mitigated in advance by defining a smaller geohash precision value and, thus, choosing a larger area, accepting a lower accuracy of the overall

dataset. Furthermore, applying filter lists on specific sensitive context factors can mitigate privacy invasions [61].

In future research, this concept can gain its effectivity in combination with other techniques. Since social media data include more and more images and videos today, pattern recognition can help detect relevant posts for disaster situations [62]. This can contribute to more precise post cardinalities and, therefore, help VOSTs to improve the groundwork for their analysis. More advanced example implementations with other than the spatial facet being the key for HLL sets could demonstrate the full potential of this technology.

## 6. Conclusions

In this paper, we have introduced a method to prevent *collateral data* when storing real-time social media data streams for analytic purposes. Our method proposes the usage of HyperLogLog (HLL), a cardinality estimation algorithm. While this technology has been applied in many contexts for the purpose of performance improvements, we newly introduced it as a method to prevent unnecessary data retention and thus protect privacy.

We embedded the method in a disaster management scenario, in which virtual operation support teams define certain terms to be monitored for occurrence and get alerted at a certain threshold. This scenario shows the usefulness of our method exemplarily for any data analysis scenario, wherein results are based only on statistical values and there is no necessity to refer to individual items in the dataset.

However, the HLL algorithm does not protect from proving existence of individual items in the set, if external knowledge is applied. The concept prevents from gaining access to previously unknown individual items. Limitations further apply for exploratory data analysis on the stored data, since the only information stored is the occurrence of a post in a certain area. This limitation is intentional for the sake of preventing unnecessary data retention of social media data and the risk of abuse, loss, or public exposure of data that were unnecessary to gather in the first place.

**Author Contributions:** Investigation, Concept, Methods, Implementation: Marc Löchner; Funding acquisition, Supervision: Dirk Burghardt. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded in the Priority Program "Volunteered Geographic Information: Interpretation, visualization and Social Computing" (SPP 1894) by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, 273827070).

**Data Availability Statement:** VGIsink proof-of-concept implementation is available at https://gitlab.vgiscience.org/ml/vgisink (accessed on 18 January 2023).

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Olteanu, A.; Castillo, C.; Diaz, F.; Vieweg, S. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1–4 June 2014.
2. Deutsche Welle. German Police under Fire for Misuse of COVID Contact Tracing App. 2022. Available online: https://p.dw.com/p/45P8H (accessed on 18 January 2023).
3. Twitter Developer Platform. Academic Research Access. Available online: https://developer.twitter.com/en/products/twitter-api/academic-research (accessed on 18 January 2023).
4. La Cava, L.; Greco, S.; Tagarelli, A. Understanding the growth of the Fediverse through the lens of Mastodon. *Appl. Netw. Sci.* **2021**, *6*, 1–35. [CrossRef]
5. Castillo, C. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*; Cambridge University Press: Cambridge, UK, 2016.
6. Fiedrich, F.; Fathi, R. Humanitäre Hilfe und Konzepte der digitalen Hilfeleistung. In *Sicherheitskritische Mensch-Computer-Interaktion*; Springer: Berlin, Germany, 2021; pp. 539–558.

7. Palen, L.; Anderson, J.; Bica, M.; Castillos, C.; Crowley, J.; Díaz, P.; Finn, M.; Grace, R.; Hughes, A.; Imran, M.; et al. Crisis Informatics: Human-Centered Research on Tech & Crises. Available online: https://hal.science/hal-02781763 (accessed on 18 January 2023).
8. Khan, Q.; Kalbus, E.; Zaki, N.; Mohamed, M.M. Utilization of social media in floods assessment using data mining techniques. *PLoS ONE* **2022**, *17*, 267079. [CrossRef] [PubMed]
9. Barker, J.; Macleod, C. Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. *Environ. Model. Softw.* **2019**, *115*, 213–227. [CrossRef]
10. de Bruijn, J.A.; de Moel, H.; Jongman, B.; de Ruiter, M.C.; Wagemaker, J.; Aerts, J.C. A global database of historic and real-time flood events based on social media. *Sci. Data* **2019**, *6*, 311. [CrossRef] [PubMed]
11. Fathi, R.; Thom, D.; Koch, S.; Ertl, T.; Fiedrich, F. VOST: A case study in voluntary digital participation for collaborative emergency management. *Inf. Process. Manag.* **2019**, *57*. [CrossRef]
12. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977; Volume 2.
13. Rojszczak, M. The uncertain future of data retention laws in the EU: Is a legislative reset possible? *Comput. Law Secur. Rev.* **2021**, *41*, 105572. [CrossRef]
14. Rucz, M.; Kloosterboer, S. Data Retention Revisited. 2020. Available online: https://edri.org/our-work/launch-of-data-retention-revisited-booklet/ (accessed on 18 January 2023).
15. Wright, D.N.; Demetres, M.R.; Mages, K.C.; DeRosa, A.P.; Jedlicka, C.; Stribling, J.C.; Baltich Nelson, B.; Delgado, D. How Long Should We Keep Data? An Evidence-Based Recommendation for Data Retention Using Institutional Meta-Analyses. 2020. Available online: https://hdl.handle.net/1813/70499 (accessed on 18 January 2023).
16. Miller, V. *Understanding Digital Culture*; SAGE Publications Limited: London, UK, 2020; pp. 145–146.
17. Guillou, C.; Portner, C. Data Retention—More Than Meets the Eye. 2020. Available online: https://www.theprivacyhacker.com/2020/12/data-retention/, (accessed on 18 January 2023).
18. Stieglitz, S.; Mirbabaie, M.; Ross, B.; Neuberger, C. Social media analytics–Challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manag.* **2018**, *39*, 156–168. [CrossRef]
19. Wang, Z.; Ye, X. Social media analytics for natural disaster management. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 49–72. [CrossRef]
20. Blanchette, J.F.; Johnson, D.G. Data retention and the panoptic society: The social benefits of forgetfulness. *Inf. Soc.* **2002**, *18*, 33–45. [CrossRef]
21. Berghel, H. Malice domestic: The Cambridge analytica dystopia. *Computer* **2018**, *51*, 84–89. [CrossRef]
22. Manokha, I. Surveillance, panopticism, and self-discipline in the digital age. *Surveill. Soc.* **2018**, *16*, 219–237. [CrossRef]
23. Büchi, M.; Festic, N.; Latzer, M. The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda. *Big Data Soc.* **2022**, *9*. [CrossRef]
24. Leetaru, K. The Era of Precision Mapping of Social Media Is Coming to an End. 2019. Available online: https://www.forbes.com/sites/kalevleetaru/2019/03/06/the-era-of-precision-mapping-of-social-media-is-coming-to-an-end/ (accessed on 18 January 2023).
25. Wilson, S. The Era of Antisocial Social Media. 2020. Available online: https://hbr.org/2020/02/the-era-of-antisocial-social-media (accessed on 18 January 2023).
26. Kuner, C.; Marelli, M., Data Analytics and Big Data. In *Handbook On Data Protection In Humanitarian Action*; International Committee of the Red Cross: Geneva, Switzerland, 2020; pp. 92–111.
27. Morton, G. *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing*; International Business Machines Co. Ltd.: Hamilton, ON, Canada, 1966.
28. Finkel, R.A.; Bentley, J.L. Quad trees a data structure for retrieval on composite keys. *Acta Inform.* **1974**, *4*, 1–9. [CrossRef]
29. Purss, M.; Gibb, R.; Samavati, F. *Discrete Global Grid Systems Abstract Specification*; Open Geospatial Consortium: Rockville, MD, USA, 2017.
30. Google. Addresses for Everyone. Available online: https://plus.codes (accessed on 18 January 2023).
31. Fischer, M. WOEID Is Deprecated. 2022. Available online: https://wiki.openstreetmap.org/w/index.php?title=Key:woeid&oldid=2367887 (accessed on 18 January 2023).
32. Dwork, C. Differential Privacy: A Survey of Results. In *Proceedings of the Theory and Applications of Models of Computation*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 1–19.
33. Victor, N.; Lopez, D.; Abawajy, J.H. Privacy models for big data: A survey. *Int. J. Big Data Intell.* **2016**, *3*, 61–75. [CrossRef]
34. Ciriani, V.; De Capitani di Vimercati, S.; Foresti, S.; Jajodia, S.; Paraboschi, S.; Samarati, P. Fragmentation and encryption to enforce privacy in data storage. In Proceedings of the European Symposium on Research in Computer Security, Dresden, Germany, 24–26 September 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 171–186.
35. Pottier, R.; Menaud, J.M. Privacy-aware Data Storage in Cloud Computing. In Proceedings of the CLOSER, Porto, Portugal, 24–26 April 2017.
36. Han, S.; Han, K.; Zhang, S. A Data Sharing Protocol to Minimize Security and Privacy Risks of Cloud Storage in Big Data Era. *IEEE Access* **2019**, *7*, 60290–60298. [CrossRef]
37. Meli, M.; McNiece, M.R.; Reaves, B. How Bad Can It Git? Characterizing Secret Leakage in Public GitHub Repositories. In Proceedings of the NDSS, San Diego, CA, USA, 24–27 February 2019.
38. Bloch-Wehba, H. Process without procedure: National security letters and First Amendment rights. *Suffolk UL Rev.* **2016**, *49*, 367.

39. Flajolet, P.; Fusy, E.; Gandouet, O.; Meunier, F. HyperLogLog: The analysis of a near-optimal cardinality estimation algorithm. *Discret. Math. Theor. Comput. Sci.* **2007**. [CrossRef]
40. Desfontaines, D.; Lochbihler, A.; Basin, D. Cardinality estimators do not preserve privacy. *Proc. Priv. Enhancing Technol.* **2019**, *2019*, 26–46. [CrossRef]
41. Ertl, O. New Cardinality Estimation Methods for HyperLogLog Sketches. *CoRR* **2017**, arXiv:1706.07290.
42. Reviriego, P.; Ting, D. Security of HyperLogLog (HLL) Cardinality Estimation: Vulnerabilities and Protection. *IEEE Commun. Lett.* **2020**, *24*, 976–980. [CrossRef]
43. Paterson, K.G.; Raynal, M. HyperLogLog: Exponentially Bad in Adversarial Settings. In Proceedings of the 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), Genoa, Italy, 6–10 June 2022; pp. 154–170. . [CrossRef]
44. Yu, Y.W.; Weber, G.M. HyperMinHash: Jaccard index sketching in LogLog space. *CoRR* **2017**, arXiv:1710.08436,
45. Löchner, M.; Fathi, R.; Schmid, D.; Dunkel, A.; Burghardt, D.; Fiedrich, F.; Koch, S. Case Study on Privacy-Aware Social Media Data Processing in Disaster Management. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 709. [CrossRef]
46. Dunkel, A.; Löchner, M.; Burghardt, D. Privacy-Aware Visualization of Volunteered Geographic Information (VGI) to Analyze Spatial Activity: A Benchmark Implementation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 607. [CrossRef]
47. Löchner, M.; Dunkel, A.; Burghardt, D. Protecting privacy using HyperLogLog to process data from Location Based Social Networks. In Proceedings of the LESSON 2019—1st International Workshop on Legal and Ethical Issues in Crowdsourced Geographic Information, Zurich, Switzerland, 8 October 2019.
48. Jiang, B. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *Prof. Geogr.* **2013**, *65*, 482–494. [CrossRef]
49. Löchner, M. VGIsink. Available online: https://gitlab.vgiscience.de/ml/vgisink (accessed on 18 January 2023).
50. Richardson, L.; Amundsen, M.; Ruby, S. *RESTful Web APIs: Services for a Changing World*; O'Reilly Media, Inc.: Newton, MA, USA, 2013.
51. Twitter Developer Platform. Available online: https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream (accessed on 18 January 2023).
52. Citus Data. Available online: https://github.com/citusdata/postgresql-hll (accessed on 18 January 2023).
53. PostGIS. Available online: https://postgis.net (accessed on 18 January 2023).
54. GeoJSON. Available online: https://geojson.org (accessed on 18 January 2023).
55. Leaflet. Available online: https://leafletjs.com (accessed on 18 January 2023).
56. Xu, J. Topic Modeling with LSA, PLSA, LDA & lda2Vec. 2018. Available online: https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 (accessed on 18 January 2023).
57. Dunkel, A.; Andrienko, G.; Andrienko, N.; Burghardt, D.; Hauthal, E.; Purves, R. A conceptual framework for studying collective reactions to events in location-based social media. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 780–804. [CrossRef]
58. Keßler, C.; McKenzie, G. A geoprivacy manifesto. *Trans. GIS* **2018**, *22*, 3–19. [CrossRef]
59. Prometheus. Available online: https://prometheus.io (accessed on 18 January 2023).
60. InfluxData. Available online: https://www.influxdata.com (accessed on 18 January 2023).
61. Burghardt, D.; Dunkel, A. Ethical Analysis of Geosocial Data to Balance Social and Individual Interests. In Proceedings of the AutoCarto 2022 International Research Symposium on cartography and GIScience, Redlands, CA, USA, 2–4 November 2022.
62. Barz, B.; Schröter, K.; Kra, A.C.; Denzler, J. Finding Relevant Flood Images on Twitter using Content-based Filters. In Proceedings of the ICPR Workshop on Machine Learning Advances Environmental Science, Online 10–15 January 2021; pp. 5–14. [CrossRef]