International Journal of Geo-Information

MDPI

Article

Case Study on Privacy-Aware Social Media Data Processing in Disaster Management

Marc Löchner [1,*], Ramian Fathi [2], David '-1' Schmid [3], Alexander Dunkel [1], Dirk Burghardt [1], Frank Fiedrich [2] and Steffen Koch [3]

1 Institute of Cartography, Technische Universität Dresden, Helmholtzstr. 10, 01062 Dresden, Germany; alexander.dunkel@tu-dresden.de (A.D.); dirk.burghardt@tu-dresden.de (D.B.)
2 Institute for Public Safety and Emergency Management, Bergische Universität Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany; fathi@uni-wuppertal.de (R.F.); fiedrich@uni-wuppertal.de (F.F.)
3 Institute for Visualization and Interactive Systems, Universität Stuttgart, Universitätstraße 38, 70569 Stuttgart, Germany; david-1.schmid@vis.uni-stuttgart.de (D.S.); steffen.koch@vis.uni-stuttgart.de (S.K.)
* Correspondence: marc.loechner@tu-dresden.de

Received: 30 September 2020; Accepted: 17 November 2020; Published: 26 November 2020

Abstract: Social media data is heavily used to analyze and evaluate situations in times of disasters, and derive decisions for action from it. In these critical situations, it is not surprising that privacy is often considered a secondary problem. In order to prevent subsequent abuse, theft or public exposure of collected datasets, however, protecting the privacy of social media users is crucial. Avoiding unnecessary data retention is an important question that is currently largely unsolved. There are a number of technical approaches available, but their deployment in disaster management is either impractical or requires special adaption, limiting its utility. In this case study, we explore the deployment of a cardinality estimation algorithm called HyperLogLog into disaster management processes. It is particularly suited for this field, because it allows to stream data in a format that cannot be used for purposes other than the originally intended. We develop and conduct a focus group discussion with teams of social media analysts. We identify challenges and opportunities of working with such a privacy-enhanced social media data format and compare the process with conventional techniques. Our findings show that, with the exception of training scenarios, deploying HyperLogLog in the data acquisition process will not distract the data analysis process. Instead, several benefits, such as improved working with huge datasets, may contribute to a more widespread use and adoption of the presented technique, which provides a basis for a better integration of privacy considerations in disaster management.

Keywords: disaster management; virtual operations support teams; privacy; data retention; hyperloglog; focus group discussion

1. Introduction

Social media services are almost always available, so that in times of crises people can access them as established and interactive communication resources. The Corona pandemic has even strengthened their special role: through measures such as social distancing, the use of social media services for communication and information dissemination is growing fast [1]. Numerous disasters in the past have demonstrated how important the role of social media is for crisis communication as well as for information gathering [2]. Both the people affected by a disaster and those indirectly affected use online services to obtain information about their relatives, the extent of any damage, possible further dangers or offers of help.

Those affected by a disaster are a vulnerable group. Some are dependent on receiving or sharing credible information or seeking help only through social media services like Facebook or Twitter. For example, in case of missing relatives, real names, addresses or even pictures of people could be published publicly. Personal data can also arise when affected people communicate publicly with authorities on social media platforms, view content or participate in self-help groups [3].

At the same time, large amounts of data are created on social media services during disasters, that are available to the public and can include important and relevant information for decision-makers. With the aim of processing and visualizing information from social media for decision-makers, so-called Virtual Operations Support Teams (VOST) are established. Consisting of volunteers, VOSTs analyze social media systematically and in a coordinated manner, and evaluate operation-relevant information.

Some of the VOSTs use social media analysis programs that store and analyze data from various social networks automatically. Such analysis software retrieves data from social media services e.g., through their application programming interfaces (API) and stores them in their own databases. This is usually necessary for the application to be able to access the actual data in a reasonable amount of time, not only since data analysis during a disaster is usually very time-critical.

Due to the usually time-critical nature of these operations, privacy aspects are prioritized rather low in disaster management. This is especially devastating because of the aforementioned very personal character of the collected data. In particular, we have determined *data retention* to be the problem with the highest urgency in this specific case.

Data retention "is the continued storage of an organization's data for compliance or business reasons" [4]. While storing the data is necessary for the analytics software to work properly, it is dangerous for its operator for multiple reasons. Constantly changing terms of services of social media services may evoke legal issues of their business [5]. Furthermore, storing large amounts of data opens the risk of possible abuse, theft or accidental public exposure [6].

The necessity arises for a method to store large amounts of social media data in a way that is technically reasonable on the one hand and privacy-aware on the other. Dunkel et al. [7] have developed a method to store and process data from any social media service in a privacy-aware fashion, using an algorithm called HyperLogLog. With this method, data queried from a social media service API can be stored in a local database, without the possibility to leak information about single entities afterwards. The resulting visual output of an analytics application will not differ significantly from what was created from conventionally stored data.

While in the past numerous studies have dealt with the topic of data analysis in crises [8] or the integration of VOST into the working environment of decision-makers [9], questions about privacy and VOST are not considered, yet. In this paper, we therefore develop approaches to answer the following central research question: what opportunities and challenges can be identified for VOSTs to work with privacy-enhanced data and what potential implementation barriers can be detected? Figure 1 outlines the concept in a graph, including stakeholders and the data flow.

To approach the research questions, we prepared and carried out a focus group discussion with voluntary VOST members, in which they were confronted with the topic and some sample dataset, which we extracted from social media services. We found, that VOST members are generally worried about losing important data, when it comes to privacy regulations. From the technical point of view they were rather indefinite, but since storing social media data using HyperLogLog both reduces storage space and increases processing speed significantly, they turned out to be quite open to a privacy-aware storage implementation.

In Section 2 we discover the fundamentals of our interdisciplinary research. Following, we describe the methods implemented in the data acquisition and the preparation and provision of the focus group discussion (Section 3). Afterwards, we discuss the results of the discussion and our findings (Section 4), concluding with an outlook of our further research plans on that topic (Section 5).

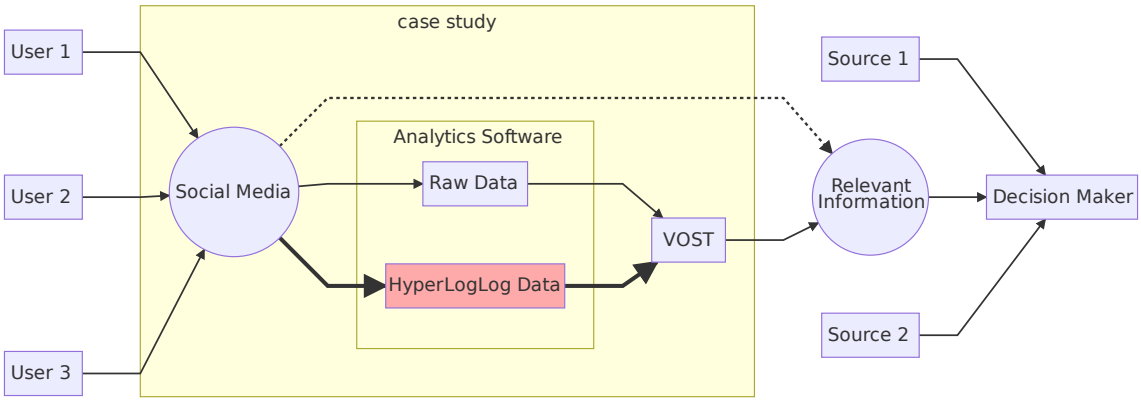

Figure 1. Users of social media services supply data, that is then parsed by analytics software, before the VOST evaluates it. Relevant information from social media is then passed to decision-makers, who found their decisions on it among other sources. The case study wraps up the part that involves social media, the analytics software and the VOST. We propose to base the analytics software on privacy-enhanced HyperLogLog data as a substitute to conventional raw data.

2. Fundamentals

In this section we explain the fundamentals and present relevant research results in the following subsection. First, we will present research that primarily addresses social media and disaster management. We examine both historical and current aspects, such as the use of online media during the Corona pandemic. Afterwards we will introduce the novel VOSTs and their tasks during a disaster. An introduction to the privacy aspect of data retention and its challenges follows. Finally, we introduce a method to store social media data in a privacy-preserving way, which we based this case study on.

2.1. Social Media and Disaster Management

As communication and collaboration platforms, social media services play an important role in disaster management, not only since the Corona pandemic. Numerous major incidents in the past have made clear that those affected, eyewitnesses or spontaneous volunteers actively use digital tools for information and communication [10]. However, there are various possible uses and behaviors, even within the different population groups. In a recent survey by Statista, 60% of the 2087 respondents (age group 18–64) state that they regularly use social networks such as Facebook, while 52% use so-called media-sharing platforms such as Instagram. In addition, 74% said they regularly use instant-messengers such as WhatsApp [11].

During crises and disasters, it can be observed that the need for information is very high, not only among the affected population but also among relatives and among other indirectly affected persons. This increased need is also reflected in the social media: In a representative survey commissioned by "Digitalverband Bitkom", 75% of users of social media services in Germany stated that they had been increasingly active in social networks during the Corona pandemic and used them more intensively. According to the survey, almost all age groups spent more time on platforms such as Facebook, Twitter, TikTok, Instagram, etc. than usual. Among the 16 to 29 year olds, the figure was as high as 86%, among the 30 to 49 year olds 82%, among the 50 to 64 year olds 74% and among the over 65 year olds 32%. At the same time, a total of 82% increasingly communicated via instant-messengers [12].

In past major incidents, such as the earthquake in Haiti in 2010, the Elbe flood in 2013 or the urban flashflood in Münster in 2014, social media helped thousands of volunteers to spontaneously network with each other and to actively participate in disaster management [13]: In Münster, hundreds of spontaneous volunteers who networked and organized themselves via social media founded their

own temporary aid organization. They participated in disaster relief in sub-groups structured parallel to the governmental emergency aid mechanisms. They carried out tasks such as easier clearing work or filling sandbags [14].

Experiences of past disasters also show that fake news can spread online within minutes, which in turn can have serious effects on psychosocial needs and government crisis management [15]. In a previous study, for example, it was possible to identify patterns and differences of false information during crisis that could be used for an automated identification of disinformation [16]. Current work such as that of Kirchner et al. examines the acceptance of social media users for approaches such as warning of misinformation [17].

A similar phenomenon is also emerging in the context of the Corona pandemic. According to Situation Report 13 of the World Health Organization [18], the outbreak and the reactions to the pandemic are accompanied by a massive "infodemic", an abundance of true and false information.

Already in 2008, Vieweg et al. [19] analyzed the use of social media in disaster relief. As a result, a new and distinct field of research developed: crisis informatics. Since then, numerous scientists from different academic perspectives have been working in this field [20–23]. Parallel to this, due to the importance of geographic data, the research area that places the use of Volunteered Geographic Information (VGI) at the forefront of scientific debate, has developed [24,25].

Since the beginning of crisis informatics research, information-technology research questions have dominated the scientific discourse. One focus has been the analysis of large amounts of data (Big Data) generated by social media [10]. Innovative research is investigating the use of artificial intelligence and methods for the machine processing during crisis [26].

Further research addresses the question of whether decision makers can actually use VGI data and whether or how they use the information provided [21]. Some studies indicate that there seems to be a significant discrepancy between the data produced by digital volunteers and the requirements of organizations involved in disaster relief [27]. Successful cooperation requires stable structures, reliable data quality and consistent forms of organization that outlast individual disaster relief operations [27].

Nevertheless, the authorities are also facing major technological challenges, especially in the area of information management. Field studies after Typhoon Haiyan in the Philippines and during the Ebola crisis (2014/2015) suggest that executives need a better understanding of their decision-making processes. It is particularly important that the needs for mission-relevant information, such as spatial data or VGI, be clearly communicated to digital volunteers [28].

In a systematic review, Reuter and Kaufhold [2] analyze 15 years of research into the use of social media in disasters, starting in 2001, the year of the attack on the New York World Trade Center. They focus on the collection of patterns of use, perception and roles that have been identified by studies in different disaster situations. Reuter and Kaufhold [29] also present derived role typologies in a matrix that differentiates between authority and citizen/public on the one hand and between real and virtual on the other. They describe that these roles and role types take either the perspective of a citizen (public) or an authority (organizational) and refer either to the real or virtual realm.

A new form of authority digital operational support are *Virtual Operations Support Teams* (VOST). They work as part of a public organization with digital volunteers in an entirely virtual environment.

2.2. Virtual Operations Support Teams

In the initial phase of crises and disasters, one of the key factors in the management of these situations is that emergency forces with the necessary skills, such as the analysis of social media for a situational awareness, are a limited resource. With the aim of processing and visualizing social media information for decision-makers, the first so-called Virtual Operations Support Team (VOST) was formed in 2011 [30]. It pursued the goal of integrating data from social media more effectively into decision-making processes. Divided into working groups, VOSTs carry out various tasks during operations:

- Social media monitoring as well as conducting data processing, filtering and assessment
- Creating and updating spatial analysis of digital maps
- Recognizing and analyzing trends and sentiment in social media
- Identifying rumors and false information
- Verifying and geolocating social media posts
- Crowdsourcing and collaborations with other VOST
- Presentation to the operational staff of the EMAs [31].

The Federal Agency for Technical Relief (Technisches Hilfswerk, THW) has been operating the first German VOST on federal level since 2017. 25 digital volunteers are deployed using innovative organizational strategies and innovative analysis tools to master the challenge of the data flood and to process situation information for decision makers. In September 2018, the Ministry of the Interior, Digitization and Migration of Baden-Württemberg established its own Virtual Operations Support Team on state level. With the time further teams followed, e.g., at the fire department Hamburg.

In a study, participating observations were used to identify organizational, procedural and technical requirements for the successful integration of VOSTs in emergency operation center structures [31]. In addition, process analyses were able to document for the first time the working methods of the VOST, such as the way information is provided to decision makers. It became clear that in many cases real-time information from publicly accessible sources contributed to an awareness of the situation and was used in time-critical decisions of the crisis management team.

The concept of these teams is also based on the fact that the helpers can work in working groups at different locations and times. Based on this high flexibility, VOST THW has already been able to manage over 25 operations. A further study shows that the public also appreciates this work. In this study, the authors show that 67% of the German population expect authorities to monitor social media. In addition, 20% of those questioned say that they use social media to search for and share information during crises and disasters [32].

2.3. The Privacy Aspect of Data Retention

Users of social media services tend to lack the awareness for the data they provide to the particular network is subject to being analyzed by third parties [33]. The daily work of VOSTs is based on data analytics software that does assemble, store and analyze data from various social networks. Storing the data in a local database is necessary for the software to be able to access it quickly, but it basically means copying of sets of data to a different place. This can as well be interpreted as *data retention*. Data retention has been defined as "the continued storage of an organization's data for compliance or business reasons" by Rouse [4].

Blanchette and Johnson [34] already addressed the disappearance of social forgetfulness as a side-effect of data retention. In terms of social networks this means that a social media user may delete their post on a network, but the post still remains existing as a copy in any dataset downloaded and stored by a third party, and the user has no chance to even take notice of this. This is especially relevant with social media data associated with crises and disasters. In certain situations, users are dependent on social networks as their only way to communicate, seek help or share credible information e.g., with public authorities [3]. Fearing their personal data being subject to theft, exposure or other abuse like mass surveillance e.g., by authorities, users may retreat from posting publicly on social media services like e.g., Twitter or Mastodon, and move to closed messaging groups in other messengers like e.g., Telegram or Matrix [35]. Miller [6] has pointed out recent incidents, where data retention has lead to theft or accidental public exposure. This underlines the significance to respect users the right to informational self-determination [36].

This would be destructive for a wide range of beneficiaries. Especially humanitarian actions rely on public availability of social media data [37]. Therefore the gradual retreat of the user base must be prevented.

2.4. Privacy-Aware Storage

Recent publications try to raise awareness for privacy aspects in humanitarian action [**?**]. They outline the necessity for suitable methods yet to be developed. VOSTs usually prioritize privacy considerations rather low due to their work being very time-critical.

A wide range of techniques to address privacy issues in big data analytics have been published, e.g., *k-Anonymity* [39] or *Differential Privacy* [40]. None of these are suitable to process huge amounts of data in a constantly updating stream, as it is the case with social media data.

A rather new technique to approach the problem of processing large datasets are *Probabilistic Data Structures* [41]. A specific example is a cardinality estimation algorithm called *HyperLogLog* [42]. Its fundamental strength is the ability to *estimate* the distinct count of a multiset (*cardinality*), stored in a data structure, that does not allow the extraction of single elements (see sample data in Figure 2). Regarding privacy protection, it is notable that without further external knowledge, it is impossible to retrieve single items out of the stored dataset [43].

Furthermore, processing data using HLL is very efficient in terms of processing time and storage space. The characteristic distinct count estimation of HLL increases the speed of processing significantly, but the result has a potential offset of about 2% [42]. When dealing with huge datasets, such as social media data, an offset of this size is negligible, because in visualizations the difference to the actual raw data is not visible.

A dataset is built upon one specific query to the original social media source, e.g., the application programming interface (API) of a social media service. Depending on the configuration of the algorithm, the dataset can only answer a certain amount of questions, that can usually be broken down to counting its elements. So HLL data can be seen as *disposable* data, since it is impossible to draw other information out of it, than originally intended. These characteristics make HLL a suitable algorithm to process social media data in a privacy-aware fashion [44].

Krumpe et al. [45] have defined a general structure to handle social media data across any social media service. Based on that structure and the HLL storage algorithm, Dunkel et al. [7] developed a method to store and process location-based social media data in a privacy-aware fashion. In this case study, we aim to investigate, whether VOSTs are able to work with data, that has been processed with this privacy-aware storage method.

3. Methods

In this section we outline the methods applied to the case study. We explain the structure of the focus group discussion and the reasoning for applied concepts. In addition, we describe how we acquired the dataset utilized as sample data in the discussion.

3.1. Focus Group Discussion

In this case study, we address the questions, what opportunities and challenges we can identify for VOSTs to work with data sets processed with the cardinality estimator HLL, and what potential implementation barriers we can detect. This issue has so far received little attention. Privacy and data protection are generally an open topic in the humanitarian aid community. Therefore, we designed a group discussion with VOST members to examine the feasibility of implementing HLL data in the VOST workflow and their sensitivity for the privacy aspect.

Participants in the discussion were members of two different VOSTs (THW and Baden-Württemberg), along with the authors of this paper, and one independent recording clerk.

Our concept intended a guideline-based discussion in order to develop approaches to answer the research questions by analyzing the discussion protocol after the discussion. The overall goal of the discussion was to document the expertise of the participants being volunteers with experience in the field.

We chose the focus group discussion method [46] because, as described in Section 2, the members of the VOSTs are very heterogeneous. They differ, for example, in their level of knowledge, working methods and working culture. As a result, there is an increased need for communication within the teams in case of operations in order to establish structures and discuss working methods. The purely virtual working method is an additional complicating factor. In focus group discussions, guideline-based and moderated sessions are encouraged so that a concrete topic can be dealt with together. The method is used in many different ways, e.g., to achieve conflict management. Nevertheless, the focus group discusson is also suitable for representing diversity of opinion and for jointly developing improvements and solutions. Because we used the established communication structures of the VOSTs, we were able to work on the specific topic using this resource-saving method.

A video conference was chosen as the venue for the discussion, using an instance of Jitsi Meet. VOST members ought to feel confortable in this environment, since they live spread across the country and virtual communication environments are their familiar place to work in the VOST (see Section 2.2).

The schedule of the discussion included a brief introduction to privacy as a concept in general and a very basic outline of the HLL data structure. The content of a sample dataset with raw and HLL-processed data was introduced along. Since the VOST members are not at all technically proficient, the introduction did not include any mathematical or computational details, but rather demonstrated the differences of raw and HLL-processed data in a tabular view (see Section 3.2).

After the introduction the VOST members were confronted with the three hypothetical scenarios and were asked to discuss their respective approach in each of them one after the other. The scenario catalog, which generally serves as a guideline for focus group discussions, was intended to establish a discussion structure and to examine the research questions. All three scenarios had the pandemic spread of the Corona virus as their basic operational situation, and only differed in the location level of operation: national, regional and local. The aim of these distinctions was the hypothesis, that different privacy challenges may occur at different levels of operation in the analysis of social media by VOST. For example, we expected that different requirements would arise at the local level than for the analysis of social media data at the federal level. Since the members of the two participating VOSTs were involved in managing such operational situations before, no detailed explanation of the scenarios was necessary.

During the discussion, however, the sample dataset only received minor attention in favor of discussing more fundamental aspects, so that the prepared scenarios did not come into action. Nevertheless, they were utilized as a guideline in which, for example, discussion questions were derived from.

The end of the discussion marked a short summary by the authors and acknowledgements to the participants.

3.2. Data Acquisition

To explain the advantages of HLL during the focus group discussion, we chose to present a sample dataset that could potentially serve as real data for a VOST operation and its members feel familiar with. So we created a dataset, that covers all German posts on Twitter containing the hashtag `#corona` from January through May 2020. Thusly, we collected `72.743.465` posts from `6.038.577` distinct users via the Twitter API. For this particular scenario, we relied on Twitter's language tagging and collected only German posts. Posts that were tagged as containing other or undefined languages were discarded. We chose to limit the collected data to only contain German posts for two reasons: (1) a German VOST team would predominantly concern with German tweets, and (2) limiting the data to contain only German posts shortened the time we spent pre-processing the data significantly. Still, pre-processing and creating the databases took about two weeks of mostly unattended computation.

We utilized the data conversion tool `lbsntransform` [47] to read the dataset into two PostgreSQL databases: one for HLL-processed data and one for the plain data, to compare against. The table in Figure 2 shows an example query to the dataset, including the structure of the HLL-processed data,

as it is stored in the database. It makes clear, that no original data can be revealed from that storage format. It further illustrates the slight difference between the actual number of posts and the *estimated* number counted by HLL (see Section 2.4).

hashtag	actual_count	hll_count	hll_data
corona	755797	763316	\x148b7f52d4d4b0c93a9696a16a4c14c42549531484a52852d083294b5456
coronavirus	614201	609185	\x148b7f4392c525a83b9886a9294a9296356b3a9473a5487a1a85ad485214
afd	391075	378847	\x148b7f521273a9074ad273a96842d08425884a5084a907629466290c3290
covid19	341884	339737	\x148b7f5252c524e639ce56a92752d4941d494a14941d283a107528c83a52
hanau	289371	277753	\x148b7f5a947320e94accb329065a108420e53a4e84a8ec521694290d5990
coronakrise	212108	215994	\x148b7f32cec4a10749cc7321075ad26424e75994631d064212d320ef31ce7
merkel	203142	201882	\x148b7f39ce74198531d0a6a4c73a94542148318c83a8e83a9273214759d6
cdu	194358	199547	\x148b7f4acc731cab4a96d518e739ca84212739cc85214841d2b69ccc72cea
berlin	183407	192788	\x148b7f425cb394e829cc9524f142106319876256949928321093bd264190
polizei	185625	185865	\x148b7f42d26521265a50641cec438e74a14861927499076a8c739d062a50

Figure 2. Most frequent descriptive metadata terms (hashtags) of posts in the sample dataset for the focus group discussion are shown in column one. The next two columns show the number of posts that contain these hashtags. Column two shows the corresponding raw count, the actual number of posts in the dataset. The equivalent HLL count, the value estimated by the HLL algorithm, is shown in column tree. The last column shows the HLL-processed data structure and emphasizes the impossibility to read original data from it.

4. Evaluation

In this section we evaluate our case study applied as a focus group discussion. First, we enumerate and explain details of the results of the discussion. A subsequent analysis of the results constitutes scientific insights, implementation examples as well as proposals for approaches and alternatives to negative results.

4.1. Findings of the Discussion

The focus group discussion examined the requirements of VOSTs in operations. Especially when VOST members get in contact with personal data of social media users, a focus on privacy aspects is crucial.

Once the participants were introduced to the privacy preserving measures, they initially were under the false assumption that more privacy would lead to less data. This would lead to them not being able to fulfill their tasks. The brief introduction into the technical features of HLL (see Section 3.1) cleared up these misunderstandings. Afterwards, participants opened up to the concept of privacy preserving aggregations, rather than using the raw data. The misconception was probable, since the VOST members were only used to working with big data. Privacy measures usually do mean that there will be less (or no more) data to work with.

However, findings of the focus group discussion encompass the disproof of our assumption that the VOSTs work is mostly based on big data analytics. In fact, the majority of efforts done by VOST members is inspecting single entities of social media data, e.g., posts, images, videos etc. As a participant already pointed out at the beginning of the discussion, an intuitive interface of the applied analysis software applied, is the far most important requirement in a time-critical work environment.

Not uncommon would be an open end investigation, where initially it is unclear to the VOST members, what they are searching for. A participant explained, it was possible that all relevant posts of a minor emergency situation, in which the VOST is activated, are first identified and then evaluated for relevance to decision-makers afterwards.

Participants of both VOSTs agreed, that one of their most important tasks is the verification of the information, that a post encompasses. This happens based on experience of each individual VOST member, and the consult of further information, usually by analyzing the meta-data of a post (e.g., time, location) and other details. It would be crucial to determine whether information in a social media post is trustworthy. Important indicators were claimed to be e.g., how long the user account

actually exists, what language(s) the account usually uses, time ranges of postings (e.g., does the user sleep sometimes) and the number of followers, likes and retweets of its postings. Another method of verifying social media posts would be to compare them with alike user names and avatars of accounts on other networks.

The participants also highlighted the fact that the scale of a disaster and the level at which an operation takes place would not have a major impact on standard work processes. They furthermore pointed out that in addition to quantitative analysis, qualitative analysis would also have to be performed manually, e.g., when assessing the relevance of a high range social media post for decision-makers. Here again, the meta-data of an account could play an important role.

An entirely different aspect, that came up during the discussion, is training of the VOST members. For exercises and practice, it is possible for VOST members to be confronted with archived datasets. This contradicts with the characteristic of HLL-processed data being *disposable*, meaning that it is not possible to draw other information than originally intended from it.

In turn, it came to light that a frequently occuring problem would be to deal with the immense size of social media data and the time to process it, accordingly. Here, using HLL-processed data can help very much, because of its characteristics to be very lightweight in storage size and fast in processing.

4.2. Analysis of the Findings

The introduction to the subject area of privacy theory and the outline of the HLL algorithm have been designed to be brief and superficial, due to the audience consisting of civil protection experts and not of information scientists. The superficiality of this introduction can be a reason, why the discussion participants showed a certain amount of skepticism about our research and a general debate about privacy issues. The discrepancy between guiding into the topic and avoiding too much technical details is apparent.

The participants expressed concerns about alleged loss or inaccessibility to important information for their work, e.g., because of privacy regulations. The volunteers explained that in disaster relief, the information content of a social media post is always more important than the personal data of a user. They are aware of the topic of privacy, but see potential challenges in its application in everyday work, specifically during a crisis. This would certainly be a problem, especially in situations where human life is in danger. In those situations, the sentiment in the team usually gets tense, and time gets critical. Other situations, e.g., tracing fake news on pandemics, do not require time-critical decisions, and allow more time to do research work.

However, according to the results of the discussion, we can indicate our key finding is that none of the findings of the discussion are opposing the proposition, the VOST could work with HLL-processed data. This can be stated, because most of the VOSTs work happens *after* any potential processing of HLL data.

Usually the VOST work starts with a search for a term, hashtag or place. The analytics software then queries the APIs of the social media services and, for technical convenience, stores the answers in a local database. Right here HLL will come into play, because it can store only the data that is really needed to answer the query.

An example query could be for the number of posts per user on a certain hashtag. The answer is a list of user IDs and the according number of posts. No further information is required, so any information on each post is unnecessary to be stored in clear text, so it can be stored in HLL. The list can then be sorted by number to identify the user having the highest number of posts. Having its' user ID, the analytics software can then query the API of the social network to get all the details about that user ID. No information about the user has been stored to the local database, and no data retention has happened.

Regarding the exercise scenario described in Section 4.1, our proposition to apply HLL data can not be applied at all. Following the mental image of disposable data, it can not be used for any other purpose. Consequently, an exercise can only lead to only one solution. Using archived HLL data from

a past situation could miss the exercise objectives. We propose the usage of publicly available example datasets with licenses that allow their utilization for purposes free of choice, like e.g., listed on the Awesome Public Datasets list [48].

As a side finding, we can declare that splitting the scenarios into three separate parts was not necessary. The size of the situation is irrelevant, since the VOST members are spread across the country and interacting mostly virtually with each other anyways.

5. Conclusions and Outlook

Privacy is an important aspect, when dealing with social media data, especially in times of crises and disaster scenarios. With HyperLogLog we proposed an appropriate technology to approach this issue and process social media data in a privacy-aware fashion. We raised the question of what opportunities and challenges can be identified for VOSTs to work with this privacy-enhanced data as a substitute for conventional datasets, and what potential implementation barriers can be detected.

In a conducted focus group discussion, we found that the underlying technology is fewer of importance to VOST members, because they hardly get in contact with it. Good interface usability and processing speed is of much larger importance, as well as comprehensive data availability. Inaccessibility or loss of data is dramatic, especially in life-depending situations. In turn, VOST members declared that dealing with social media data can be stressful due to the plain size, resulting in large storage requirements and slow processing speed. This may sound like a contradiction. However, the results of our research presented in this paper show that with using appropriate technology it is certainly possible to implement a method to provide privacy-aware infrastructure for social media data processing. But we also have to take into account that by selecting representatives from only two (German) VOSTs, other constellations of participants might produce other findings. The goal of our case study was to develop first approaches to answer our research question, we did not pursue the goal of a representativeness of the results.

These findings encourage further research on that topic. Our first subsequent step is to develop a qualitative survey featuring an interview-style experiment with VOST members. This should feature an experimental setup based on HLL-processed data, that will be presented to the participants in their familiar working environment. Following a mix of established survey methods *behavior coding* and *think aloud protocol* [46], they should explain their working steps in a supervised surrounding and express their user experience regarding limitations or barriers. We are confident, that such an experiment will confirm our findings of this work and lead to an actual implementation in common social media analytics software.

Author Contributions: Conceptualization: Marc Löchner, Ramian Fathi, David '-1' Schmid and Alexander Dunkel; Data curation: David '-1' Schmid; Funding acquisition: Dirk Burghardt; Investigation: Marc Löchner, Ramian Fathi and David '-1' Schmid; Methodology: Marc Löchner, Ramian Fathi and David '-1' Schmid; Project administration: Marc Löchner; Resources: Ramian Fathi and David '-1' Schmid; Software: Marc Löchner and David '-1' Schmid; Supervision: Alexander Dunkel, Frank Fiedrich and Steffen Koch; Writing—original draft: Marc Löchner, Ramian Fathi and David '-1' Schmid; Writing—review and editing: Marc Löchner and Ramian Fathi. 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).

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

References

1. Koeze, E.; Popper, N. The Virus Changed the Way We Internet. Available online: https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html (accessed on 25 November 2020).
2. Reuter, C.; Kaufhold, M.-A. Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics. *J. Conting. Crisis Manag.* **2018**, *26*, 41–57. [CrossRef]

3. Kuner, C.; Marelli, M. Social Media. In *Handbook on Data Protection in Humanitarian Action*; International Committee of the Red Cross: Geneva, Switzerland, 2020; pp. 223–237.
4. Rouse, M. Data Retention. Available online: https://searchstorage.techtarget.com/definition/data-retention (accessed on 25 November 2020).
5. Fiesler, C.; Beard, N.; Keegan, B.C. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta Georgia, GA, USA, 8–11 June 2020; Volume 14, pp. 187–196.
6. Miller, V. *Understanding Digital Culture*; SAGE Publications Limited: London, UK, 2020; pp. 145–146;
7. Dunkel, A.; Löchner, M.; Burghardt, D. Privacy-Aware Visualization of Volunteered Geo-Graphic Information (Vgi) to Analyze Spatial Activity: A Benchmark Implementation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 607. [CrossRef]
8. Reuter, C.; Hughes, A.L.; Kaufhold, M.-A. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *Int. J. Hum. Comput. Interact.* **2018**, *34*, 280–294. [CrossRef]
9. St Denis, L.A.; Hughes, A.L.; Palen, L. Trial by fire: The deployment of trusted digital volunteers in the 2011 shadow lake fire. In Proceedings of the 9th International ISCRAM Conference, Vancouver, BC, Canada, 22–25 April 2012; pp. 1–10.
10. Castillo, C. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*, 1st ed.; Cambridge University; Cambridge University Press: New York, NY, 2016; ISBN 978-1-107-13576-5.
11. Statista Global Consumer Survey Umfrage in Deutschland zu beliebten Arten von Social Media 2020: Welche Arten von Social Media nutzen Sie regelmäßig? Available online: https://de.statista.com/prognosen/999854/umfrage-in-deutschland-zu-beliebten-arten-von-social-media (accessed on 25 November 2020).
12. Bitkom e.V Social-Media-Nutzung Steigt durch Corona stark an. Available online: https://www.bitkom.org/Presse/Presseinformation/Social-Media-Nutzung-steigt-durch-Corona-stark-an (accessed on 25 November 2020).
13. Sackmann, S.; Lindner, S.; Gerstmann, S.; Betke, H. Einbindung ungebundener Helfer in die Bewältigung von Schadensereignissen. In *Sicherheitskritische Mensch-Computer-Interaktion*; Reuter, C., Ed.; Springer: Wiesbaden, Germany, 2018; pp. 529–549, ISBN 978-3-658-19523-6.
14. Fathi, R.; Rummeny, D.; Fiedrich, F. Organisation von Spontanhelfern am Beispiel des Starkregenereignisses vom 28.07. 2014 in Münster. *Notfallvorsorge* **2017**, *2*, 1–8.
15. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.-C.; Flammini, A.; Menczer, F. The Spread of Low-Credibility Content by Social Bots. *Nat. Commun.* **2018**, *9*, 4787. [CrossRef] [PubMed]
16. Starbird, K.; Maddock, J.; Orand, M.; Achterman, P.; Mason, R.M. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In *iConference 2014 Proceedings*; iSchools: Washington, DC, USA, 2014.
17. Kirchner, J.; Reuter, C. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–27. [CrossRef]
18. World Health Organization Novel Coronavirus(2019-nCoV): Situation Report 13. Available online: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf (accessed on 25 November 2020).
19. Vieweg, S.; Palen, L.; Liu, S.B.; Hughes, A.; Sutton, J. (Eds.) *Collective Intelligence in Disaster: Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shooting*; University of Colorado: Boulder, CO, USA, 2008.
20. Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; Vieweg, S. AIDR— Artificial intelligence for disaster response. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 11 April 2014; pp. 159–162.
21. Imran, M.; Castillo, C.; Diaz, F.; Vieweg, S. Processing Social Media Messages in Mass Emergency. *ACM Comput. Surv.* **2015**, *47*, 1–38. [CrossRef]
22. Leysia, P.; Sarah, V.; Jeannette, S.; Sophia, B.; Liu, A.H. Crisis Informatics: Studying Crisis in a Networked World. In Proceedings of the Third International Conference on E-Social Science, Ann Arbor, MI, USA, 4–9 October 2007.
23. Qadir, J.; Ali, A.; ur Rasool, R.; Zwitter, A.; Sathiaseelan, A.; Crowcroft, J. Crisis Analytics: Big Data-Driven Crisis Response. *J. Int. Humanit. Action* **2016**, *1*, 1–21. [CrossRef]
24. Goodchild, M.F. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]

25. Goodchild, M.F.; Glennon, J.A. Crowdsourcing Geographic Information for Disaster Response: A Research Frontier. *Int. J. Digit. Earth* **2010**, *3*, 231–241. [CrossRef]
26. Alam, F.; Ofli, F.; Imran, M. Descriptive and Visual Summaries of Disaster Events Using Artificial Intelligence Techniques: Case Studies of Hurricanes Harvey, Irma, and Maria. *Behav. Inf. Technol.* **2019**, *3*, 1–31. [CrossRef]
27. Hughes, A.; Tapia, A. Social Media in Crisis: When Professional Responders Meet Digital Volunteers. *J. Homel. Secur. Emerg. Manag.* **2015**, *12*, 203. [CrossRef]
28. Comes, T.; Vybornova, O.; Van de Walle, B. Bringing structure to the disaster data typhoon: An analysis of decision-makers' information needs in the response to haiyan. In Proceedings of the AAAI Spring Symposium, Stanford, CA, USA, 23–25 March 2015.
29. Reuter, C. (Ed.) *Sicherheitskritische Mensch-Computer-Interaktion: Interaktive Technologien und soziale Medien im Krisen- und Sicherheitsmanagement*; Springer Vieweg: Wiesbaden, Germany, 2018; ISBN 978-3-658-19523-6.
30. Fathi, R.; Schulte, Y.; Schütte, P.; Tondorf, V.; Fiedrich, F. Lageinformationen Aus Den Sozialen Netzwerken: Virtual Operations Support Teams (Vost) International Im Einsatz. *Notfallvorsorge* **2018**, *49*, 1–9.
31. 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.* **2020**, *57*, 102–174. [CrossRef]
32. Reuter, C.; Kaufhold, M.-A.; Schmid, S.; Spielhofer, T.; Hahne, A.S. The Impact of Risk Cultures: Citizens' Perception of Social Media Use in Emergencies Across Europe. *Technological Forecasting and Social Change* **2019**, *148*, 119724. [CrossRef]
33. Polous, K. Event Cartography: A New Perspective in Mapping. Ph.D. Thesis, Technische Universität München, Munich, Germany, 2016.
34. Blanchette, J.-F.; Johnson, D.G. Data Retention and the Panoptic Society: The Social Benefits of Forgetfulness. *Inf. Soc.* **2002**, *18*, 33–45. [CrossRef]
35. Leetaru, K. The Era of Precision Mapping of Social Media Is Coming to an End. Available online: https://web.archive.org/web/20191219100123/https://www.forbes.com/sites/kalevleetaru/2019/03/06/the-era-of-precision-mapping-of-social-media-is-coming-to-an-end/ (accessed on 25 November 2020).
36. Eberle, E.J. The Right to Information Self-Determination. *Utah Law Rev.* **2001**, *2001*, 965.
37. 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.
39. Samarati, P.; Sweeney, L. Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement Through Generalization and Suppression. 1998. Available online: https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf (accessed on 25 November 2020).
40. Dwork, C. Differential privacy: A survey of results. In Proceedings of the International Conference on Theory and Applications of Models of Computation, Kitakyushu, Japan, 13–16 April 2008; pp. 1–19.
41. Singh, A.; Garg, S.; Kaur, R.; Batra, S.; Kumar, N.; Zomaya, A.Y. Probabilistic Data Structures for Big Data Analytics: A Comprehensive Review. *Knowl.-Based Syst.* **2020**, *188*, 104987. [CrossRef]
42. Flajolet, P.; Fusy, É.; Gandouet, O.; Meunier, F. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. In Proceedings of the Discrete Mathematics and Theoretical Computer Science, Dijon, France, 7–12 July 2007; pp. 137–156.
43. Desfontaines, D.; Lochbihler, A.; Basin, D. Cardinality Estimators Do Not Preserve Privacy. *Proc. Priv. Enhanc. Technol.* **2019**, *2019*, 26–46. [CrossRef]
44. Löchner, M.; Dunkel, A.; Burghardt, D. Protecting Privacy Using Hyperloglog to Process Data from Location Based Social Networks. In Proceedings of the Legal Ethical factorS crowdSourced geOgraphic iNformation 2019: 1st International Workshop on Legal and Ethical Issues in Crowdsourced Geographic Information, Zürich, Switzerland, 8–9 October 2019.
45. Krumpe, F.; Dunkel, A.; Löchner, M. LBSN Structure. Available online: https://pypi.org/project/lbsnstructure/ (accessed on 25 November 2020).
46. Prüfer, P.; Rexroth, M. *Verfahren zur Evaluation von Survey-Fragen: Ein Überblick*; Social Science Open Access Repository: Mannheim, Germany, 1996.

47. Dunkel, A. Lbsntransform. Available online: https://pypi.org/project/lbsntransform/ (accessed on 25 November 2020).
48. Chen, X.; Pival, P.R.; Cirik, A.; Mohajerani, S. Many More Awesome Public Datasets. Available online: https://github.com/awesomedata/awesome-public-datasets (accessed on 25 November 2020).