The Legal and Ethical issues Workshop @LREC2022, pages 27–37
Marseille, 20 June 2022

About Migration Flows and Sentiment Analysis on Twitter Data: Building the Bridge Between Technical and Legal approaches to data protection

Thilo Gottschalk, Francesca Pichierri
FIZ Karlsruhe - Leibniz-Institute for Information Infrastructure
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen

{francesca.pichierri, thilo.gottschalk }@fiz-karlsruhe.de

Abstract

Sentiment analysis has always been an important driver of political decisions and campaigns across all fields. Novel technologies allow automatizing analysis of sentiments on a big scale and hence provide allegedly more accurate outcomes. With user numbers in the billions and their increasingly important role in societal discussions, social media platforms become a glaring data source for these types of analysis. Due to its public availability, the relative ease of access and the sheer amount of available data, the Twitter API has become a particularly important source to researchers and data analysts alike. Despite the evident value of these data sources, the analysis of such data comes with legal, ethical and societal risks that should be taken into consideration when analysing data from Twitter. This paper describes these risks along the technical processing pipeline and proposes related mitigation measures.

Keywords: sentiment analysis, data protection, privacy

1. Introduction

Social media data are commonly processed for analysing and predicting social phenomena. They are relatively easy to obtain, cheap and contain a lot of valuable and diverse information - ranging from factual to subjective (Pereira-Kohatsu et al., 2019; Ligthart et al., 2021). Tweets are particularly popular among researchers due to their accessibility, actuality and ease of processing (Ligthart et al., 2021; Goritz et al., 2019). One particular field that shows strong interest in the use of such data is the field of migration studies and border security as can be observed by public funding directed towards research in this area[1], by research activities in general (Carammia et al., 2022), as well as by Frontex strategical-analysis documents[2] and public tenders (Frontex, 2019). In the field of migration studies, Twitter data analysis is considered very useful for a series of purposes such as measuring and predicting migration flows, providing necessary support to vulnerable groups/migrants/refugees, assessing the integration of migrants in destination countries or evaluating public opinion towards migration (Righi, 2019; Mijatović, 2021). The importance of such approaches was most recently highlighted in the context of the Ukraine war where social media intelligence (SOCINT) played an important role (Engelhaupt, 2022).

Despite the practical and analytical advantages, the processing of Twitter data can raise concerns regarding the right to data protection and privacy of Twitter users as well as affected third parties. Linkage of different datasets can produce a clearer picture of global migration flows but also raise risks for unwanted and inappropriate negative societal effects, e.g. for migrants and refugees.

Given these contrasting effects, it is crucial to design and implement analytical models and approaches in a way that balance technical and data protection requirements without undermining compliance with the legal framework, such as the General Data Protection Regulation (GDPR), nor the purposes of the data analysis. This alignment proves to be difficult especially for data scientists with no deeper understanding of the legal frameworks they conduct their work in. At the same time, legal experts often lack sufficient understanding of the technical approaches and the possible risks linked to them. This often results either in over-regulation or in non-compliance of the processing.

Where risks and potential negative consequences towards users are identified at an early stage, it is possible to adopt mitigation measures to address these risks and foster compliance with the data protection by design and data protection by default principle, enshrined in Article 25 GDPR. That being said, compelling approaches require interdisciplinary efforts involving legal experts as well as developers and data scientists to find a common language that is intelligible to all parties and to break down the knowledge barriers between different fields of expertise.

On these grounds, this paper aims to provide a foundation for structured approaches towards privacy preservation in the analysis of Twitter data and aims to build a bridge between technical and legal data protection approaches in Twitter data driven sentiment analysis. In-

[1] See e.g. the EU project ITFLOWS: `https://www.itflows.eu/`; the project METICOS:`https://meticos-project.eu/`; and EFFECTOR:`https://www.effector-project.eu/`.

[2] `https://frontex.europa.eu/we-know/situational-awareness-and-monitoring/strategic-analysis/`.

spired by the analytical work conducted in the project ITFLOWS (IT tools and methods for managing migrations FLOWS)[3], this paper focuses on the use of Twitter data to detect risks of tensions related to migration and it directs the attention towards sentiment analysis performed on such data. On the context of migration research we explain legal and societal impacts of sentiment analysis on Twitter data, providing insights and guidance on common risks of technical approaches and how to mitigate them. While a variety of approaches have been discussed and proposed to ensure privacy of data subjects in data analysis, these approaches often either refer to structured data or neglect the technical pipeline of such approaches.

Such discussions are hence often difficult to follow for technical personnel, are not always suitable for unstructured social media content (such as textual Twitter data) and do not reflect the technical reality when processing personal data. In line with this, it can be observed that many existing research papers and approaches pose considerable risks to the data subject (e.g. simple re-identification, annotation) and correlating liability risks to the data controller. A very common problem are publicly available annotated datasets that contain not only analytical outcomes but the Tweet-ID as well. This, one the one hand, makes the research reproducible. On the other hand, it also allows easy identification of the Twitter user together with potentially sensitive information (e.g. sentiments towards specific topics). Such data can easily be used to identify and target members of certain groups for political advertisements, making the abstract data protection risk a concrete problem.[4] Neither the researchers, nor the affected data subjects are usually aware of this risk. With this paper we strive to highlight such risks and mitigation measures linked to the technical steps that typically compose the *sentiment analysis*.

While it is impossible to cover all existing analytical methods and techniques in the field of sentiment analysis, the paper aims to provide a starting point that can be used to develop a compelling *privacy aware* approach on a case-by-case basis for data driven sentiment analysis. It provides contextual and technical guidance and applicable substance to the more generic legal requirements imposed by the GDPR. The proposed structure can be used to validate research/processing approaches and thereby aims to foster legal and ethical sustainability within and beyond research approaches.

The paper starts by presenting the necessary background data protection concepts as laid down by the GDPR (Section 2). A data scientist in the role of controller needs to take proactive actions to ensure and demonstrate compliance with the obligations set by the GDPR, from the beginning to the end of processing. Therefore, particular attention will be directed towards the explanation of accountability-based mechanisms, such as the principles of data protection by design and by default under Article 25 GDPR and Data Protection Impact Assessments (DPIAs) under Article 35 GDPR. Secondly, the analysis moves towards a description of the general technical approach of Sentiment Analysis in the context of Twitter data (Section 3). Sentiment analysis can be conducted on Tweets by means of different techniques (Thakkar and Patel, 2015; Saberi and Saad, 2017). While there is no standard solution to the processing of social media data for the purpose of sentiment analysis, there are multiple (linked) processing steps that tend to play an important role and which regularly appear in one or another form in sentiment analysis methodologies. Such technical steps provide the structure for the analysis conducted in this paper. In principle, each step in the technical pipeline can raise risks but also provides a potential leverage point to mitigate overall risks (see Section 4) to data protection and privacy.

The paper hence addresses legal researchers and data scientists alike. We aim to provide understandable technical insights to legal scholars and foster the understanding of the data protection implications and technical solutions for data scientists/developers. The analysis invites data scientists to rethink their technical processes in favor of a privacy preserving perspective, provides a source for acknowledged and feasible mitigation measures and aims to strengthen the legal and technical capability to communicate the respective needs by providing recommendations for the identified analytical/processing steps.

2. Data Protection obligations

The GDPR imposes obligations onto data scientists when processing information through which it is possible to identify a natural person (personal data).

Under the GDPR, processing means 'any operation or set of operations which is performed on personal data or on sets of personal data' (Art. 4 (2) GDPR). It includes collection, recording, storage, alteration, use, dissemination, combination or erasure - in principle, the definition includes any possible operation that could be performed on personal data.

Data scientists could process personal data in the role of controllers when determining, alone or jointly with others, the purposes and the means of the processing (Art. 4 (7) GDPR) or they could do it in the role of processors when processing personal data on behalf of the controller(s). Data controllers are the primary bear-

[3]The goal of the ITFLOWS project is to provide accurate predictions and adequate management solutions of migration flows in the European Union. The project develops precise models which lay the foundation of the EUMigraTool (EMT), a software platform that will provide to relevant stakeholders a set of tools enabling simulations and predictions. The EMT has two main functions: predicting migration flows and detecting risks of tensions related to migration, `https://www.itflows.eu/`.

[4]Most of publicly available Twitter datasets contain TweetIDs, we hence refrain from referencing a specific one here.

ers of the obligations set by the regulation towards the person whose data is processed (data subjects), while data processors faces a limited number of obligations (see e.g. Art. 30 and Art. 32 GDPR). By nature, social media data usually have at least some relation to the publishing user. Contrary to wide believe (especially among data scientists), public availability must not be mistaken for consent to be freely used in any other context.

2.1. Data Protection Principles

When processing personal data, both controllers and processors need to comply with the general data protection principles listed in Art. 5 GDPR.

2.1.1. Lawfulness

Processing must be lawful, i.e. it must respect all applicable legal requirements. The core conditions for processing to be lawful are listed in Art. 6 GDPR. Processing is lawful only if and to the extent that at least one of the conditions listed applies, such as, for example, consent of the data subject, necessity for the performance of a task carried out in the public interest, necessity for the purposes of the legitimate interests pursued by the controller or a third party, if such interests are not overridden by the interests or rights and freedoms of the data subject (Art. 6 (1) GDPR). Furthermore, in Art. 9 the GDPR identifies some types of personal data which are particularly sensitive and merit enhanced protection, such as those revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, those concerning health or sexual life or sexual orientation, genetic data and biometric data processed for the purpose of uniquely identifying a natural person. The processing of such sensitive data, in principle, is prohibited pursuant to Art. 9 (1) GDPR, unless one of the exemptions in Art. 9 (2) GDPR applies. Exemptions include situations where the data subject has given consent, where processing relates to personal data which are manifestly made public by the data subject; where processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Art. 89 (1) GDPR. The latter often arguably provides a legal foundation for the processing, however, the Article particularly requires that the processing must be subject to appropriate safeguards, in accordance with the GDPR, for the rights and freedoms of the data subjects.

2.1.2. Fairness

Processing must be fair and conducted in an ethical manner. For example, data must not be obtained through unfair means, such as by deceiving data subjects or by acting without their knowledge.

2.1.3. Transparency

Processing must be transparent to the data subject concerned. The controller is obliged to take any appropriate measures to keep data subjects informed regarding the processing of their personal data before and during the processing activities and also in regard to a request of access. Information should be easily accessible and easy to understand. Elements concerning content and quality of the information duty are subject of Art. 12-15 GDPR.

2.1.4. Purpose limitation

Data must be collected for specified, explicit and legitimate purposes and not further processed in a manner incompatible with those purposes. The purpose of processing must be determined before processing is started and it must be unambiguous and clearly expressed. Furthermore, the purpose must be balanced between the rights and interests of the controller and the ones of the data subject. Each new purpose for data processing which is incompatible with the initial one must have its own specific legal basis. Exceptions to this rule are considered for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes (Art. 5 (1) (b) GDPR), with the application of appropriate safeguards (Art. 6 (4) GDPR; Recital 50 GDPR).

2.1.5. Data Minimization

Processed personal data must be adequate, relevant and limited to what is necessary in relation to the purposes specified. Instead of a "process everything approach", such principle promotes a selective method which begins prior to collection and concerns not only the quantity but also the quality of personal data. It also requires to ensure that the period for which personal data are stored is limited to a strict minimum (Recital 29 GDPR). This principle also remains applicable under the research exemptions as laid down in Art. 89 GDPR.

2.1.6. Accuracy

Personal data must be accurate and kept up to date. In every processing activity, the controller must take every reasonable step to ensure respect to this principle. All inaccurate personal data should be erased or rectified without delay.

2.1.7. Storage Limitation

Personal data must be stored in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed. Personal data must be deleted or anonymised as soon as they are no longer needed. Controllers are encouraged to establish time limits for erasure or for a periodic review (Recital 39). The storage limitation principle permits the storage of personal data for longer periods if it is processed exclusively for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Art. 89 (1) GDPR and it is subject to implementation of the appropriate technical and organizational measures in order to safeguard the rights and freedoms of individuals.

2.1.8. Integrity and Confidentiality

Personal data must be processed in a way that ensures its appropriate security, integrity and confidentiality, including protection against unauthorized or unlawful processing, against accidental loss, damage or destruction. To ensure this, appropriate technical and organizational measures need to be implemented. Chapter IV of the GDPR (from Art. 24 to Art. 43) provides guidance to controllers and processors on how to adequately fulfill such principle.

2.1.9. Accountability

The controller is responsible for, and must be able to demonstrate compliance with, all the previous principles listed. Such requirement is further developed in Art. 24 GDPR.

2.2. Ensuring compliance with the obligations

In addition to the data protection principles listed above, the controller has to implement mechanisms to comply with the rights of the data subject laid down in Chapter III of the GDPR (from Art. 12 to Art. 23 GDPR).

The controller needs to take proactive actions to ensure and demonstrate compliance with the obligations set by the GDPR, from the beginning to the end of processing. To this purpose, appropriate and effective technical and organizational measures must be implemented; additionally, they need to be reviewed and updated where deemed necessary (Art. 24 GDPR). The determination of the measures to be taken depends on the processing being carried out, the types of data processed and the level of risk to data subjects. Ways to facilitate compliance include ensuring Data Protection by Design and by Default (Art. 25 GDPR) and conducting a Data Protection Impact Assessment (Art. 35 GDPR).

2.2.1. Data Protection by Design and by Default

Addressing data protection issues at a very early stage, when designing and setting up processing strategies and activities is crucial. Data Protection by Design means embedding data protection principles and safeguards in the design and development of data processing models, therefore ensuring protection of privacy-related interests right from the start (when the means for processing are determined). This requires that, conceptually, the relevant measures are defined prior to the system being set up, rather than implementing measures ex post.

By making data protection an important element of the core functionality of an analytical model, the controller is facilitated in ensuring a privacy compliant solution, allowing the processing to meet data protection requirements and ensure protection of data subjects´rights. Art. 25 (1) GDPR requires the implementation of appropriate technical and organisational measures (e.g. pseudonymisation) taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks to data subjects.

Data Protection by Default requires the controller to ensure that, by default, only personal data which are necessary to achieve a specific purpose of the processing are processed. This applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. This would also mean for example avoiding using technical solutions that collect more personal data than are strictly necessary for a specific functionality or which do not ensure confidentiality.

Processors are not obliged to assist controller with data protection by design and default obligations (unlike with security measures under Art. 32 GDPR). However, controllers must select processors that provide sufficient guarantees to meet the GDPR´s obligations (Art. 28 GDPR). Breach of Art. 25 GDPR may result in the imposition of sanctions (Art. 83 (4) GDPR).

2.2.2. Data Protection Impact Assessment

The Data Protection Impact Assessment, DPIA, is a requirement provided by Art. 35 GDPR. The DPIA´s objective is to evaluate the impact of the planned processing activities on the protection of personal data and it must be carried out by the controller prior to processing. The assessment should contain at least:

(a) a systematic description of the data processing and the purposes of the processing and where applicable – the legitimate interests of the controller;

(b) an assessment of the necessity and proportionality of the data processing on the basis of the specified purpose;

(c) an assessment of the risks to the data subjects rights and freedoms (e.g. likelihood and severity)[5]

(d) measures proposed to address these risks, including safeguards, security measures, mechanisms to ensure personal data protection and to demonstrate compliance with the Regulation (see Article 35 (7) GDPR).

The DPIA is both an "accountability measure" as well as a "warning system" (Kuner et al., 2020, p. 699). The outcome of the assessment is helpful in the determination of "appropriate measures" to be carried out in order to demonstrate compliance with data protection principles and obligations.[6] Through the DPIA, risks and potential negative consequences of processing activities to data subjects can be identified at an early

[5] According to Recital 76 GDPR, "The likelihood and severity of the risk to the rights and freedoms of the data subject should be determined by reference to the nature, scope, context and purposes of the processing. Risk should be evaluated on the basis of an objective assessment by which it is established whether data processing operations involve a risk or a high risk".

[6] See Recital 84 GDPR.

stage. The controller can evaluate and propose mitigation measures to address the risks identified and significantly limit the probability of negative outcomes. This identification and evaluation exercise supports compliance with the data protection by design and default principle. Although the regulation specifies that the DPIA must be carried out before the processing starts, it is advisable controllers see the DPIA as a process where data processing operations, risks and measures put in place are managed and reviewed continuously.

The DPIA is required in cases where a data processing operation is likely to result in high-risk to the rights and freedoms of individuals, in particular if it makes use of new technologies. According to Recital 75 GDPR, the risk "may result from personal data processing which could lead to physical, material or non-material damage". For example, the processing may give rise to discrimination, identity theft, financial loss, damage to reputation, economic or social disadvantage. Art. 35 (3) GDPR provides a non-exhaustive lists of processing likely to result in high risk (Datenschutzkonferenz (DSK), 2018). These include cases where special categories of data, e.g. information on racial or ethnic origin, political opinions, religious or philosophical belief, is being processed on a large scale[7]. Recital 75 GDPR mentions cases where personal aspects are evaluated in order to create or use personal profiles, e.g. when aspects concerning personal preferences, behaviour, location, movement are analysed or predicted; it also mentions cases where personal data of vulnerable natural persons are processed. Data protection authorities in part provide examples of processing scenarios that by default do or do not result in a DPIA obligation (Brink and Wolff, 2021, *Hansen*, Art. 35 Rn. 13). In addition, the WP29 DPIA guidelines lay out a list of criteria which can be taken into account when establishing whether processing activities are "likely to result in high risk" (Article 29 Working Party, 2017, p. 8-11).

DPIAs are not mandatory for all data processing activities as the obligation is tied to the existence of a likely high risk. However, it has been pointed out that, in practice, a controller needs to always conduct a preliminary assessment of the processing activities to identify whether the latter are likely to result in a high risk and therefore in need of a DPIA (Kuner et al., 2020, p. 671). Furthermore, in general it may be prudent to conduct DPIAs, whether or not the high-risk standard is met, or in doubt of it. The DPIA is, in fact, a very useful tool that helps controllers to comply with data protection law, ensure best practices and minimize liability (Article 29 Working Party, 2017, p. 9).

There is no specific DPIA template, although there are some valuable suggested formats that can be taken into consideration (e.g. (Information Commissioners Office (ICO), 2017; Commission Nationale de l'Informatique et des Libertés (CNIL),)). Controllers may also develop their own templates. When carrying out a DPIA, controllers can seek the advice of the Data Protection Advisor where designated.[8] Furthermore, they do not necessarily need to conduct the assessment on their own but can also outsource the DPIA to third parties (Brink and Wolff, 2021, *Hansen*, Art. 35 Rn. 11).

[7] Article 35 (3) (b) GDPR.

3. Sentiment Analysis

Processing in Sentiment Analysis, especially on social media data, often results in high risk for the data subjects. In the case of Twitter, every single Tweet is at least related to the author of the Tweet and can be related to an undefined number of natural persons. As sentiment analysis is often linked to sensitive topics, there is a high risk that special categories of data (Art. 9 GDPR) are processed. In consequence, it becomes particularly important to mitigate data protection risks in all processing steps.

3.1. Definition of Sentiment Analysis

"Sentiment Analysis is the review of written or other forms of communication or qualitative data to determine a quantifiable and comparable measure of some form of feeling in the communication or data" (Peslak, 2017, p. 38). In other words, it is a computational study of people´s affective states in relation to a particular entity, such as a topic or event, which aims to create "actionable knowledge" (Ligthart et al., 2021, p. 4998). Sentiment analysis is a complex process that usually consists of numerous tasks, such as subjectivity classification and sentiment orientation (Saberi and Saad, 2017; Ligthart et al., 2021). Information related to sentiments or opinions concerning a specific topic are mined from a word, sentence or document (level of analysis) and, in a simple approach, sentiments are classified into positive (denoting the state of happiness, satisfaction etc.), negative (denoting the state of discontent, anger etc.) and in a few cases also into neutral (when no sentiment has been detected). Factual information are discarded as SA is directed towards subjective sentences (Saberi and Saad, 2017) but can play a role in the interpretation of the analysis.

Datasets for SA are usually user-generated textual content. To this end, social media data proved to be a particularly valuable source of data as it is highly subjective and full of informal language, i.e. textual content. In this context, Tweets are particular popular as they are easily obtained, contain real-time/recent[9] information on topics and they have a similar format (Ligthart et al., 2021).

[8] Article 35 (2) GDPR; see also Recital 84 GDPR.

[9] the currentness of the data depends on the TwitterAPI. E.g. the Firehose API provides real-time access to all tweets, standard access limits access to a certain time-window, research access is limited to historical data but not to a specific time-window.

3.2. Technical Approach to Sentiment Analysis

Sentiment Analysis is an umbrella term and can be conducted by means of different techniques and approaches. In the context of Twitter, the analysis is usually based on textual data of tweets. To this end, the technical approaches usually rely on various forms of natural language processing paired with additional methods aligned in a processing pipeline (Thakkar and Patel, 2015; Saberi and Saad, 2017; Ligthart et al., 2021). The design of a processing pipeline, i.e. the linked methods and concepts, to conduct sentiment analysis on the available Twitter data can be manifold. While there is no standard order to the processing of social media data for the purpose of sentiment analysis, there are multiple (linked) processing steps that tend to play an important role and regularly appear in one or another form in sentiment analysis. The order and the relevance of the steps is driven by various factors such as the available data, purpose of the analysis, expertise and available tools. We hence describe common processing steps in a logical, yet not obligatory, order. Such steps can be broken down into

1. *Source Identification* (3.2.1),
2. *Data Collection* (3.2.2),
3. *Data Cleansing* (3.2.3),
4. *Data Analysis* (3.2.4).

For each of these steps there are uncountable options to conduct the necessary tasks. A task can be conducted by means of complex machine learning (ML) based approaches (which can be unsupervised, supervised and semi-supervised), semantic-based analysis or hybrid/combined approaches but also by more simplistic processing approaches (e.g. counting Tweets). For the purposes of this paper we will depict selected exemplary approaches in each step in order to shed light on common approaches and - in combination with Section 4 - provide insights on the legal implications, risks and potential mitigation measures that need to be considered when planning to conduct data driven sentiment analysis.

For example, Topic Modelling can be used for purposes of data cleansing (Section 3.2.3) as well as for the actual analysis of data with regard to sentiments. The goal is hence not to investigate one specific approach in a specific context but rather to emphasize the importance to acknowledge where and why a certain approach was chosen. Further research on specific legal implications in a given context would be desirable but are out of the scope of this paper.

3.2.1. Source identification

Prior to the actual analysis, a feasible data source and appropriate data selection/extraction methods need to be defined. Both steps are of utmost importance and need to be aligned with the purpose of the research. To do so, the research question must be clearly defined and narrow enough to provide a proper definition of the processing purpose. Where available, different sources should be taken into consideration and there should be reasonable explanation for the chosen data source. For Twitter data, valid sources can be the Twitter API itself, but also (pre-processed) sources[10] such as Knowledge Bases (e.g. TweetsKB (Fafalios et al., 2018) or MigrationsKB (Chen et al., 2021b)).

Recommendation: Sources should be identified not only by accessibility and volume but also lawfulness of their creation, legal (e.g. Terms of Use, domestic law) and practical limitations (e.g. difficult data cleansing; re-identification possibilities). A reasoning for the chosen source and the weighing of interest has to be provided.

3.2.2. Data Collection

Rule-based content extraction is the most straight forward approach to limit extraction of data to relevant topics. All major social media platform allow access to user-generated content through APIs and thereby (usually) allow extraction of data based on keywords (i.e. only a certain topic, such as *migration*) or metadata (i.e. only data from June-August) or other queries (Calisir and Brambilla, 2018, p. 116). However, as already pointed out by Calisir and Brambilla (2018), such traditional approach often lacks specificity and results in noisy outcomes, two shortcomings that could, among other, clash with the GDPR demands regarding data quality, (sufficient) data minimization and/or purpose limitation.

Twitter provides an API to access tweets in a structured manner. The Twitter API provides access to *Tweets, Users, Direct Messages, Lists, Trends, Media, Places* and currently consists of two versions and multiple tiers (Twitter API v2; Standard v1.1; Premium v1.1 Enterprise). The tiers provide different but overlapping features. From a data protection perspective, the used tier should be driven by the underlying question/purpose of the processing. The respective API functionality is further limited by context of processing (currently *Standard Project, Academic Research Project*). In the context of academic research, Twitter allows access to the full-archive endpoints (Tweet counts, Tweet search). However, in the EU framework the access to the archives is additionally governed by the GDPR and the granted contractual access by Twitter must not be mistaken for a hall pass to process Twitter data without limits.

In addition to general legal requirements (e.g. GPDR), the use of Twitter data is governed by private agreements such as the Twitter Developer Policy[11] in which Twitter imposed their own privacy and data protec-

[10] (e.g. http://www.sentiment140.com/)

[11] `https://developer.twitter.com/en/developer-terms/policy`.

tion principles on the users. In addition to the Developer Policy, Twitter provides governance through various rule sets (e.g. *Automation Rules, Display Requirements, API restricted Uses Rules, Twitter Rules, Twitter Brand Resources, Periscope Community Guidelines, Periscope Trademark Guidelines, Batch compliance*) which are not examined within this article but constitute private agreements between the data user and Twitter. These private agreements are partly reflecting GDPR requirements but are also logically governed by Twitter's economic interests and liability considerations rather than fundamental rights aspects and shift liability towards the end user. To a great extent the ToU limit access to Twitter data to what is lawful under the applicable legal frameworks, but sometime also excess legal requirements. However, - in contrast to popular believe - the ToU, provision of data through the Twitter API or even signed contracts *do not* automatically result in the lawfulness of the processing to the data but solely limit Twitter´s liability through shifting responsibility to the developers. Lawfulness from a data protection perspective is, hence, solely driven by the factual circumstances of the processing as laid down in the GDPR. Twitter *expects* the developers to comply with the national and international regulations on their own, although the Twitter API and access restrictions are designed to support developers in their endeavour to act lawful.

Beyond the contractual agreement between Twitter and the developers/controllers, additional rules are imposed on the data controller through generally applicable legal instruments such as the GDPR. These *general* obligations are often referred to in the contractual agreements and compliance with them is subject to contractual obligations. However, the obligation to comply with general requirements does not stem from contracts but rather from the law itself (i.e. these obligations exist independently, with or without reference in the ToU). Failure to comply with general data protection obligations results in an unlawful infringement of the data subjects rights to data protection and privacy and a contractual breach in relation to Twitter.

Recommendation: The collection of data should be conducted with the lowest privacy impact possible for the specific approach. Targeted collection is preferable to ex-post data cleansing. To this end, the respective API documentations (where available) should be checked and a reasoning for the chosen approach in the light of the purposes of the analysis has to be provided. This encompasses limitations in the volume (e.g. timeframes) as well as content-limitations (e.g. exclude Hashtags, Usernames)

3.2.3. Data Cleansing

For most data analysts, data cleansing is a technically relevant step to make the analysis efficient. However, it should also be seen as a way to ensure compliance with the principle of data minimisation as laid down in Art. 5 (1) (c) GDPR and which requires removal of any personal data that is not required for the analysis.

The chosen cleansing approach often depends on the used libraries (e.g. in python NLTK, re, spaCy, gensim, scikit, TensorFlow, or MITIE, text2vec, Moses in C++). Libraries can be described as a toolbox that can be used in various scenarios with multiple different tools (or methods) that can be applied alone or in conjunction depending on the specific needs. In consequence, there are uncountable different approaches to conduct the primary analysis. To reduce impacts on fundamental rights and interests it is important to choose a) the correct libraries b) the right tools (c.f. 4). Some libraries (e.g. NLTK) provide specific guidance/documentation how to use Twitter data (Bird and Tan,), however, due to the myriads of applications of NLP, general discussions or descriptions of anonymization methodologies are usually not provided. GDPR compliant pre-processing/preparation of data hence remains in the hands of the data scientists using these tools. Prior to any analysis, sensitive pieces of text need to be identified and then masked via suppression or generalization approaches (Hassan et al., 2021, p. 1).

Which information has to be filtered out depends on the specific purpose of the processing. Accordingly, the *why* and *how* of the chosen cleaning methods should be clearly explainable. Data controllers should be able to provide proper reasoning as to why certain tools or methodologies for data cleansing are used. For example, Tweets can filtered/cleaned using tools such as *Presidio* (Microsoft, 2022) to identify and exclude personal data from text or erasing geo-data, by relying on filtering of specific keywords or through application of topic modelling approaches. In the given example, a valid reason could be that, while extracting tweets based on the keyword "migrant", tweets concerning "migrant birds" (far away from the topic of migration flows and border control) can also be included, whereas detection of Tweets through topic modelling becomes more precise (Chen et al., 2021b) (Chen et al., 2021a).

In addition to the principle of data minimisation, the principle of data accuracy (Art. 5 (1) (d) GDPR) can impose further requirements on the data controller. To this end, Twitter provides an API endpoint to check offline data against the status of the Twitter database. This endpoint is intended to help controllers/developers to identify if the Twitter content (and hence the underlying intent of the Twitter user) may have changed. To achieve this, the respective dataset with each line containing either the Tweet IDs or the user IDs can be uploaded to the endpoint. Twitter will internally compare the dataset against the internal Twitter data and provide the developer with a set of JSON objects relating to a specific Tweet providing information if the *Tweet or account was deleted, Tweet or account was deactivated, geo data was removed, account is protected or*

account was suspended. This compliance check should be conducted on a regular basis and prior to any major analytical approaches. Failure to test the own dataset against Twitter data can result in inaccurate data and infringed not only Twitter´s ToU but more importantly the GDPR principle of data accuracy laid down in Article 5 (1) (d) GDPR.

The foundation of the sentiment analysis can usually broken down to a NLP task. To this end, the processing of data consists of *Tokenization* and data cleaning. Tokenization breaks down textual data into smaller 'pieces' (i.e. tokens). Tokens are often single words but can also be hashtags, emoticons, multiple words or other information embedded in textual data received from the Twitter data. In an additional step, these special characters and stop words are usually removed from the dataset to make processing more efficient and then accessible (e.g. in an array) for further analytical steps in the processing pipeline. To this end, it should be specified which methodology was used for Tokenization and why. In addition, it needs to be acknowledged that Tokenization becomes more difficult for some languages. This can direct the focus towards English tweets due to relative ease of Tokenization of English language with "out-of-the-box" solutions. As a consequence Tokenization can shift sentiment analysis to certain user groups which can generate unforeseen bias in the outputs and should be properly reflected in the interpretation of SA outcomes.

Pursuant to Art. 5 (1) (d) GDPR, the data controller needs to ensure that the stored data reflects the user intent and the current state of content on Twitter. In consequence, an additional pre-processing/cleansing step should be layered on top of the traditional cleansing steps for NLP. To ease this, Twitter provides data controller with a *Batch compliance*[12] procedure, which is unfortunately not very openly communicated and hence often unknown to data users.

Recommendation: Presumed compliance with the ToU does **not** automatically constitute legal compliance under the applicable law - especially in international contexts. The driving factor when designing data processing approaches should be the applicable legislation (e.g. GDPR and national specifications). Existing approaches to foster data accuracy, such as Twitter´s Batch Compliance procedure should be used unless more efficient approaches are available. Tokenization should not only be seen as a necessary preparatory step for the analysis but also as a step to remove personal information from the dataset.

3.2.4. Analysis

As mentioned above, the analysis is strongly dependent on the purpose of the processing, the available skills and tools. Accordingly, analytical approaches cannot be comprehensively covered within a single paper but require contextual analysis and research. Regularly used approaches in the context of Twitter data, for example, make use of *Metadata extraction, Topic Modelling, Entity linking* – to name a few. In discussions between legal and technical personnel the focus often lies on the analysis only, neglecting impacts of the preparatory steps described above.

Sentiment analysis is a text categorization task with the goal to extract a positive or negative orientation that text expresses toward some object based on features of the data (i.e. in the unstructured Tweet text). In principle, is a classification task that - in its simplest form - can be described as multi step approach based on classification of words in a sentence (positive +1; negative -1; neutral 0) that results in a calculation of the final score of the sentence to detect the sentiment to an object. An object in this sense can be anything (e.g. a movie, a book, migrants or a political party). In a simple approach, sentiment analysis can be a binary classification task that simply checks for certain words that are considered either positive (excellent, great) or negative (awful, ridiculous). However, the ruleset for such a simplistic approach has its limits when classification tasks becomes more complex. More refined and complex approaches hence become increasingly important and, for example, rely on supervised or unsupervised machine learning approaches not only linked to single words but contextual information (e.g. reflected through ngrams, topics). Usually the algorithm should learn to return a predicted class for new/unknown documents. Despite complex rulesets, outcomes will usually provide a probability that a document belongs to class (i.e. reflects a positive or negative sentiment). From a legal perspective, the key component that needs to be assessed is the underlying classifier and the used ruleset. For example, it can be distinguished between *generative* (e.g. naive Bayes) and *discriminative* (e.g. logistic regression) classifiers. Generative classifiers build a model how a class (e.g. sentiment) would generates input data (e.g. document). Provided an observation, i.e. an unknown Tweet, the classifier can identify which class would most likely generate such an observation. In the context of sentiment analysis, Topic Modelling – for example – can be useful to identify topics represented in Tweets, to detect negatively connoted Tweets. Equally, topic modelling can also be used to identify relevant tweets as part of the data cleansing (Section 3.2.3).

Topic modeling techniques can be used to extract and categorize "hidden semantic structures" in a textual data, such as a tweet; in simple terms, word frequency and patterns are detected and grouped in order to identify topics (Chen et al., 2021b). This procedure can easily be conducted over hundreds of thousands of Tweets and/or other sources. This means that instead of using a whole bag of words, the words that reflect a certain topic are identified. A topic, in this case, not nec-

[12] `https://developer.twitter.com/en/docs/twitter-api/compliance/batch-compliance/quick-start`.

essarily reflects the "human" understanding of a topic but can be used to detect subjective information such as opinions, attitudes, and feelings expressed in text (Lin and He, 2009; Onan et al., 2016). From a data protection perspective, the data controller should be able to provide a valid reasoning why a specific approach was chosen from the available options (e.g.e.g. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Biterm Topic Model (BTM) (Yan et al., 2013), Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Non Negative Matrix Factorization (NMF), Parallel Latent Dirichlet Allocation (PLDA), Pachinko Allocation Model (PAM)). Similarly, there are many different approaches for extraction and classification (e.g. unsupervised, semi-supervised, supervised techniques) and diversity can be found in datasets, area of interest and language used (Rana et al., 2016). Where classifications rulesets are generated by a machine-learning approach, their underlying concept should be subject to a legal and ethical assessment prior to the use of such approach. Depending on the training methods, the ruleset can contain information that is not interpretable by humans. If the methodology (here: ruleset) cannot be assessed by a natural person, it should be acknowledged that, in principle, it can result in unforeseen or unwanted (although mathematically correct) outcomes.

Among others, possible reasons for a decision towards a specific approach can be the efficiency of the approach, necessity of extraction of implicit or explicit topics, lawfulness of underlying datasets and their quality. With regard to the latter, a challenge can be, for example, the lack of consensus on a definition of "hate-speech" that contributes to the problem of finding reliably annotated data (Kovács et al., 2021; Zhang and Luo, 2019). This becomes particularly important where rules to determine sentiment are driven by supervised machine learning and the supervised classifier is trained on documents d that have been hand-labeled with a class c (i.e. positive or negative). For example, BERT (Bidirectional Encoder Representations from Transformers) is a machine learning model used for NLP tasks (Devlin et al., 2018) that enables processing of each token (e.g. word) of input text in the full context of all tokens before and after. In addition, models are usually pre-trained on a large corpus of text and then fine-tuned for specific task (transfer learning). Shortcomings in annotated data hence have to be considered by the data controller, especially if data can be linked to natural persons during or after the analysis (e.g. false-positive identification of hate-speech linked to an identifiable individual). In consequence, such approaches should only be used on properly cleaned data sets to avoid any linkage with personal data. In addition, various methodologies may raise a general risk to reinforce biases or misunderstandings. The data controller should hence be sufficiently skilled to compare different approaches, evaluate the pros and cons not only from a technical but also from a legal and ethical perspective. This is also the case where the analysis relies on publicly available libraries. Such libraries may benefit from community inputs – however, their output and/or correctness should not be assumed and the risk of perpetuating biased or faulty concepts underlying the library should be assessed and acknowledged.
When analysis outputs are intended to be reused or published (e.g. research articles) or made available in data collections (e.g. Knowledge Bases or Knowledge Graphs (Fafalios et al., 2018; Mendes et al., 2012)) that enable further analysis and/or research with the data, the aforementioned risks are multiplied. Such activities can hence result in an infringement of data subjects rights to privacy and data protection and generate liability concerns for the data controller (i.e. the data scientist). In consequence, it is recommended enforce access restrictions to such data – even though it may seem undesirable in research context at first glance.

Recommendation: The analytical approaches are usually driven by the underlying research question, respective expertise and practical limitations (e.g. computational power). Nevertheless, the shortcomings of the chosen approaches and implications for further analysis should be made transparent and also be included where data sets or results are shared/published.

4. Legal, Ethical and Societal Implications

As shown above, sentiment analysis comes with various risks and challenges that can be addressed on multiple levels. In general, the data controller should be aware that mitigation measures can and have to be applied during the planning, pre-processing, analysis and subsequent use (e.g. publication) of data. In all of these phases, technical and organizational mitigation measures have to be taken into consideration. If mitigation of risks is not possible in an early phase, the corresponding risk has to be addressed and mitigated later in the process. This could, for example, mean that the analysis itself is lawfully possible but research outcomes could not be published because the data could not be cleaned properly in the pre-processing (or later on). The illicit publication of personal data results in liability risks for the data controller, but also generates broader ethical and societal risks (e.g. misuse of research outcomes for political advertisements).
Beyond the technical mitigation measures in the respective steps, it is the task of the data controller to transparently communicate remaining risks and what further measures might be taken to reduce legal, ethical and societal risks especially when data is reused. Especially in the area of research, it should be carefully considered who will have access to the method itself and/or the outputs of the data. Indiscriminate access to generated datasets bear a high risk for misinterpretation and/or misuse especially if the shortcomings and mitigation measures implemented in the pre-processing and/or analysis are not properly addressed. Mitigation

of these risks could, for example, require the usage of a data trustee as envisioned in the novel Data Governance Act (DGA) to enable third parties access to the (research) data. To date, research in sentiment analysis is widely focused on the "efficiency" of a method in comparison to other approaches. While this approach is compelling from a mere research perspective it would be desirable to accompany these research aspects with legal, ethical and societal risks and how they can be addressed within the respective pipeline/approach. Such discussions, currently only take place in a very limited scope and data scientists often see such risks as hindering rather than guiding for their own research.

(Hassan et al., 2021, p. 3) point out that the detection of personal information in unstructured textual data suffers from severe limitations as current approaches a) often fail to detect identifying phrases, b) detect natural entities (NE) that should not be suppressed from the analysis (e.g. references to countries) and c) only detect NE that they have been trained to use. While these shortcomings are true, these approaches still provide a feasible anonymization solution in some contexts. However, the data controller needs to be aware of the respective shortcomings and should be able to provide a reasoning why a certain library, tool or approach was used.

During and after the analysis it is important to acknowledge the risk of error manifestation, depending on the analytical method. Sentiment analysis approaches that are based on topic modelling hence have to be subject to regular evaluation and usage of such approaches in operational contexts should be subject to human review. Correctness should not be taken as granted (e.g. specific terms have been identified as hate-speech/negative in 2020 but the same term has a different connotation a few years later). Published outcomes and/or datasets (e.g. Knowledge Bases) should hence properly depict the underlying methodology as well as the measures to ensure privacy of the data subjects as well as correctness of the outcomes. Such information should be manifested not only in accompanying publications but rather linked with data directly in form of metadata. Where personal information remains in the data (e.g. because data cleansing is not possible), it is particularly relevant to acknowledge and address risks in the accuracy of analysis. Risks can be connected, for example, to the "simple" fact that people express sentiments in complex ways (e.g. the use of irony, sarcasm, humor)(Saberi and Saad, 2017, p. 1664), and often texts contain slangs, abbreviations, typos, incomplete information and implicit language which challenge basic classification (Ligthart et al., 2021, p. 5031). At the same time, the categorization of sentiments into two or three groups (positive, negative and neutral) inevitably oversimplifies the complexity of sentiments´ affective qualities and this has to be always kept in mind. Depending on the social context, application of sentiment analysis further bears a specific risk for misinterpretation. To this end, the analysis as well as the interpretation of sentiment analysis should acknowledge different meanings/sentiments depending on the region and context of use (e.g. the word *cunt* has very different meanings reaching from very negative to positive in GB, AUS while in Canada the very same word is seen exclusively negative and offending). Where outcomes are published - e.g in a Knowledge Base -, they should reflect these social and contextual differences (Kovács et al., 2021). Given the technical difficulty in representing societal differences at least the user/researcher needs to be aware of them and address them when building upon a KB.

In a broader picture, opinion and sentiment mining can contribute to a "chilling effect" or "self-censorship effect" that should be countered with transparent processing approaches, lawful processing (i.e. proper data cleansing) as well as open discussion about risks and shortcomings of the used approaches (Manokha, 2018; Kennedy, 2012).

All of the aforementioned steps and mitigation measures require further research both on the legal and technical level. To this end, it would be desirable to further foster interdisciplinary efforts in both realms.

5. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 882986.

6. Bibliographical References

Article 29 Working Party. (2017). Guidelines on data protection impact assessment (dpia) and determining whether processing is "likely to result in high risk" for the purposes of regulation 2016/679, wp 248 rev.01, brussels, 4 october 2017.

Bird, S. and Tan, L.). Nltk - twitter howto.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brink, S. and Wolff, H. A. (2021). BeckOK DatenschutzR.

Calisir, E. and Brambilla, M. (2018). The problem of data cleaning for knowledge extraction from social media. In Cesare Pautasso, et al., editors, *Current Trends in Web Engineering*, pages 115–125, Cham. Springer International Publishing.

Carammia, M., Iacus, S. M., and Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, 12(1):1–16.

Chen, Y., Gesese, G. A., Sack, H., and Alam, M. (2021a). Temporal evolution of the migration-related topics on social media.

Chen, Y., Sack, H., and Alam, M. (2021b). Migrationskb: A knowledge base of public attitudes towards migrations and their driving factors. *CoRR*, abs/2108.07593.

Commission Nationale de l'Informatique et des Libertés (CNIL).). Privacy impact assessment.

Datenschutzkonferenz (DSK). (2018). List of processing activities for which a DPIA is to be carried out.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Engelhaupt, E. (2022). Social media crackdowns during the war in ukraine make the internet less global. *ScienceNews*.

Fafalios, P., Iosifidis, V., Ntoutsi, E., and Dietze, S. (2018). Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference*, pages 177–190. Springer.

Frontex. (2019). Service Contract for the Provision of Social Media Analysis Services Concerning Irregular Migration Trends and Forecasts (as part of Prewarning Mechanism) - Frontex/OP/534/2019/DT. accessed 12-October-2021.

Goritz, A., Kolleck, N., and Jörgens, H. (2019). *Analyzing Twitter data: Advantages and challenges in the study of UN climate negotiations*. SAGE Publications Ltd.

Hassan, F., Sanchez, D., and Domingo-Ferrer, J. (2021). Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Information Commissioners Office (ICO). (2017). What is a dpia?

Kennedy, H. (2012). Perspectives on sentiment analysis. *Journal of Broadcasting & Electronic Media*, 56(4):435–450.

Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media. *SN Computer Science*, 2(2):1–15.

Kuner, C., Bygrave, L., Docksey, C., and Drechsler, L. (2020). *The EU General Data Protection Regulation: A Commentary*. Oxford University Press. Available at: https://global. oup. com/academic ....

Ligthart, A., Catal, C., and Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, pages 1–57.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Manokha, I. (2018). Surveillance, panopticism, and self-discipline in the digital age. *Surveillance & Society*, 16(2):219–237.

Mendes, P. N., Jakob, M., and Bizer, C. (2012). Dbpedia: A multilingual cross-domain knowledge base.

Microsoft. (2022). Presidio - data protection and anonymization api.

Mijatović, D. (2021). COE - Commissioner for Human Rights: A distress call for human rights. The widening gap in migrant protection in the Mediterranean".

Onan, A., Korukoglu, S., and Bulut, H. (2016). Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7(1):101–119.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Peslak, A. (2017). Sentiment analysis and opinion mining: current state of the art and review of google and yahoo search engines' privacy policies. *Journal of Information Systems Applied Research*, 10(3):38.

Rana, T. A., Cheah, Y.-N., and Letchmunan, S. (2016). Topic modeling in sentiment analysis: A systematic review. *Journal of ICT Research & Applications*, 10(1).

Righi, A. (2019). Assessing migration through social media: a review. *Mathematical Population Studies*, 26(2):80–91.

Saberi, B. and Saad, S. (2017). Sentiment analysis or opinion mining: a review. *International Journal of Advanced Science Engineering Information Technology*, 7:1660–1667.

Thakkar, H. and Patel, D. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. *arXiv preprint arXiv:1512.01043*.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.