ICIC Express Letters
Volume **13**, Number **7**, July **2019**

 ISSN 1881-803X
pp. **609–616**

COMPLIANCE: A BIG DATA APPROACH WITH LAW AND BUSINESS DOMAIN EXPERT ASSESSMENT

Daniel Moritz Marutschke[1], David Marutschke[2]
and Hans-Peter Marutschke[3]

[1]College of Information Science and Engineering
Ritsumeikan University
1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
moritz@fc.ritsumei.ac.jp

[2]Faculty of Business Administration
Soka University
1-236 Tangi-machi, Hachioji-shi, Tokyo 192-8577, Japan
mdavid@soka.ac.jp

[3]Doshisha Law School
Doshisha University
Kamidachiuri-sagaru, Karasuma-dori, Kamigyo-ku, Kyoto 602-0023, Japan
hmarutsc@mail.doshisha.ac.jp

Received December 2018; accepted March 2019

Abstract. *With the General Data Protection Regulation in operation since May 25, 2018, awareness in the professional world and the rest of society differs by a large margin. This research was prompted by what domain experts presumed as skewed information in the public domain. To investigate one of the major sources of current information dissemination for the public domain and as of 2018, this research describes the method of text-mining keyword-filtered messages from the social media website Twitter and how the findings are evaluated from domain experts in business and law. The combination of using big data acquisition, text mining, and domain experts collaboration was expected to shed some light on such issues. The authors find that related terms, news items, and discussions can be identified. The main discussion is in addressing absence of relevant education in compliance topics. This paper details the methodology and analysis to gain insight into the difference in perceived compliance and expert assessment.*
Keywords: Big data mining, Twitter, Compliance, GDPR

1. **Introduction.** Organizational compliance has long been part of business culture, at least on paper. How these hold up when scrutinized is often a different story. Compliance affects many facets of society, with a newly sparked urgency in the business and legal world due to the General Data Protection Regulation, which recently – as of May 25, 2018 – became operative. Legal and business experts are widely aware of the importance and implications of the new EU regulations and were inclined to investigate the perception of them on a more extensive audience.

Less reliable polling and popular media outlets were a trigger for the authors to start a collaboration in investigating the main narratives in news outlets, sourced by social media platforms. It is unsurprising the lack of knowledge displayed on most outlets as the topic is complex not only from a business administrative, but also legal perspective.

In combination with big data mining of social media platform Twitter, an Information and Communication Technology (ICT) approach was chosen in a cross-domain effort to examine social perception and related articles and reports. Twitter was expected to emphasize oversimplified messages due to the limited amount of characters that can be

DOI: 10.24507/icicel.13.07.609

used[1]. The authors, however, were interested in the co-occurring articles and their analytical depth.

The rest of this paper is structured into five sections. A business perspective and insight into values from research related to organization is given in Section 2. The following Section 3 provides background information on the General Data Protection Regulation in respect to compliance. Section 4 details the acquisition of the data set, analysis, and methodology used in this research. Results and related discussions are provided in Section 5. Concluding remarks and future work are summarized in Section 6.

2. **Compliance from the Business Domain Perspective.**

Scholars from various fields have investigated approaches on how organizations can nurture a culture of compliance and how they can design and implement specific corporate ethics programs [1, 2, 3, 4, 5, 6, 7]. However, digital transformation is taking place throughout many organizations at a fast pace, which makes it increasingly challenging to establish new or adjust existing processes that meet the requirements of a dynamically changing environment. Recent discussions about data privacy, cybercrime, and cyberwar have made clear that pressure on management to ensure that employees follow ethical and legal rules has never been so high. The authors try to address the discrepancy between the fast-pace digital evolution and business and law adaption by a joint analysis of ICT, big data mining, and expert evaluation.

A culture of compliance is a fundamental requirement for businesses to succeed in the long term, which is not created simply by a written list of company rules and annual training sessions [2, 8]. Also ethics play an important role in business activities, as it is generally accepted knowledge in business management, that companies perform best, if they have clear ethical values and behave in accordance with those values, involving all stakeholders [9, 10].

Research suggests that Gallup's meta-analysis of employee engagement for example shows that business units with highly engaged employees have 28% less internal theft or shrinkage than their bottom-quartile counterparts[2] [11].

Nevertheless, many companies still follow a reactive and fragmented approach, where security initiatives lack continuity and consistency. As a result, many risks and hidden threats are uncovered too late or, in the worst case, remain undetected [12]. The digital transformation now requires companies to rapidly collect, measure, and analyze compliance data to predict and act fast on possible threats. Gallup refers to a variety of methodologies such as external benchmarking, anonymous reporting, pulse surveys, and focus groups.

Taking the given background into consideration, this paper uses a big data approach with the help of Twitter, one of the largest social media platforms to disseminate trending information, to find opportunities for investigation in the dawn of the General Data Protection Regulation.

3. **Background on Compliance and the General Data Protection Regulation.**

Since provisionally agreed upon in 2015, the General Data Protection Regulation (GDPR) of the EU was scheduled to finalize in May 2016 [13, 14]. The mandatory compliance was set with a two-year waiting period for businesses and organizations operating in or with the EU to get ready to comply with the new regulations on May 25, 2018.

[1]One message (*Tweet*) was limited to 140 multibyte characters, the limit has since September of 2017 been extended to 280 characters, except for multibyte glyphs, which count stays the same. `https://twitter.com/`

[2]Refer to Gallup online resources at `http://news.gallup.com/businessjournal/190352/managing-employee-risk-requires-culture-compliance.aspx`. Last accessed October 20, 2018

There have been many discussions related to business, individual, and research implications since, especially from the data collection standpoint [14, 15, 16].

The GDPR has been designed mainly to protect the Personally Identifiable Information (PII) on EU residents and will replace existing local data protection laws with a modern regulation designed for the data and Internet age and backed up with some big punitive measures for non-compliance.

According to the European Commission, the GDPR helps so that "people have more control over their personal data" and "businesses benefit from a level playing field"[3].

The European Commission digital library PDF describes the following main changes[4]: policies described in clear and straightforward language; need for affirmative consent; need for clear information on data transfer; data collection only for well-defined purposes; need to inform on automated algorithmic decision making; need to inform without delayed in case of data breach; possibility for the user to move data to competing services; user access to a copy of personal data; user option to have personal data erased. These practices are assured to be enforced by 28 data protection agencies as the European Data Protection Board with authority to penalize businesses in breach.

In order to understand the complexity and practical importance of the new law, its main content can be summarized as follows.

The GDPR is mainly focused around consent, legitimate use and other aspects of data protection. Although data security occupies little of the text it does have big significance with new stricter, more specific, obligations on both data processors and controllers. There are no specific controls but instead both controllers and processors are required to "implement appropriate technical and organizational measures" (Art. 24 Sec. 1 GDPR). This is qualified by referencing "the state of the art and the costs of implementation" (Preliminary Note 83 GDPR) and "the nature, scope, context, and purposes of the processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons" (Preliminary Note 74 GDPR).

Although some of these issues are already covered by existing data protection laws, the GDPR goes further and suggests what kinds of security controls might be considered "appropriate to the risk" (Art. 32 Sec. 1 GDPR), including:

- The pseudonymization (this can be viewed as reversible anonymization) and encryption of personal data.
- The ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services.
- The ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident.
- A process for regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing.

To demonstrate compliance with the GDPR the controller or processor should "maintain records of processing activities" (Preliminary Note 82 GDPR). Precondition to control PII within an organization/company, the locations and systems where PII might be found have to be discovered and documented first. In most organizations collections of so-called *dark data* exist, which are data hidden from the known or formal infrastructure – these databases can vary from small data stores on individual user's PCs to large database applications which are not being managed as part of the core infrastructure – and may leak outside of the organization into third parties. Technologies exist to locate and document where PII might exist. These are typically called *Data Discovery* tools – some of which

[3]The official European Commission website can be accessed via: `https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en`. Last accessed October 20, 2018

[4]`https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf`. Last accessed October 20, 2018

are configured to find particularly sensitive types of data such as credit-card numbers, racial terms, personal identifiers and data patterns. These tools could search through an entire connected infrastructure – networks, PC's servers and even mobile devices and catalogue all the data discovered. Importantly this can be a basis for a *PII Data Asset Register* which will become a vital asset to meet any form of data compliance [17].

The law will be enforced by 28 data protection agencies as the European Data Protection Board and respective agencies in the member states with authority to penalize businesses in breach. Severe penalties will be imposed on organizations in breach or non-compliance of GDPR. They can be fined up to 4% of annual global turnover or €20 Million (whichever is greater being the maximum fine), if they do not have sufficient customer consent to process data or violating the core of Privacy by Design concepts.

Compliance with the GDPR is therefore of essential importance and a vital interest for all private and public organizations, as mentioned above, not only in the EU [18, 19].

4. **Data Set and Analysis Methodology.** To get a timely insight into the changes of attitude and the propagation of information regarding the General Data Protection Regulation, the authors reasoned to investigate the most prevalent real-time news feed and social media platform Twitter. It was deemed important to collect data directly in dawn of the new regulations taking effect. This investigation was, however, only possible by collaboration of experts in fields of law, business, and information and communication technology. This section focuses on the data collection aspect and data mining to be examined from a legal and business viewpoint.

To investigate news outlets, opinions, and other factors, the social media platform Twitter was selected to gather a data set of 14,326 messages, including metadata. Founded by Jack Dorsay in March 2006, Twitter has an active user base of 330 million per month. With 500 million messages sent per day, over 80% are written from mobile devices. The amount of data makes it one of the prominent providers in big data and text mining.

The methodology used in this paper follows five major steps:

1) Data acquisition based on compliance-related key terms
2) Data pre-processing
3) Automated word ranking using frequency tables
4) Extracting associated URLs in the original text message based on previous high-content value words
5) Evaluation and interpretation of the articles and reports by domain experts

The following paragraphs go into details on each of the above steps.

Data acquisition based on compliance-related key terms. For effective text mining purposes, a Python program was written to collect and process the messages gathered from the Twitter platform. For keyword search and language selection of Twitter Status Objects, the `TwitterSearch` package was used. English and German Tweets were searched separately by filtering for language first, as included by the `TwitterSearch` package. The following keywords were targeted for each of the languages respectively: "`compliance`", "`GDPR`", and "`DSGVO`", the latter being the German equivalent to the GDPR, standing for *Datenschutz-Grundverordnung*.

The data set acquisition was set to cover a timespan of one month, within which the new GDPR was introduced. Using the `Twitter Rest API`, Twitter Status Objects in `JSON` (JavaScript Object Notation[5], the native file format of a Twitter Status Object) were collected and processed. Each Twitter Status Object represents a status update (message, or *Tweet*) of an account. Messages are bundled with metadata, including the following:

[5]Refer to `www.json.org` for detailed implementation and usage.

- user information, including the following:
  - user ID, use name, date of account setup, mission statement, follower count, geo tag information, language settings, over ten individual design choices for the visual profile, time zone, and more
- the Twitter Status Object ID
- the actual message content as UTF-8 encoded text
- location data, including the following:
  - profile location, name of country and city, location of the tweet, GPS coordinates (listed with decreasing likelihood the user has turned these features on)
- further metadata to track conversations of a connected message (*in-reply-to* relation),
  - the name, ID, status, etc.

Data pre-processing. After the data was gathered, it was processed to facilitate further analysis in the following steps:

1) Extract text-only content from each Twitter Status Object
2) Pre-processing: using the Python implementation of the Natural Language Toolkit (`NLTK`), stop-words and irrelevant tags or code fragments – e.g., "`RT`" or "`https`" – were removed
3) User handles starting with "`@`" and usually in the beginning of a message were removed using regular expressions

Following the text extraction and pre-processing, the remaining words with high content value are listed into a frequency table. This has similar implications to word co-occurrence and is used in this case to assess words commonly associated with the keyword *GDPR* and *compliance*. Filtering the original Twitter Status Objects by high content value words, the content of available URLs in these text messages was inspected by domain experts in law and business, the co-authors of this paper.

Extracting associated URLs in the original text message based on previous high-content value words. Based on the results from the previous step, URLs that were included in Twitter messages with words high on the frequency table were compiled. This was done by targeting the original Twitter Status Object's `text` segment and extracting the code fragment containing the website. Another frequency table was then built to list the websites by how often they were referred to.

Evaluation and interpretation of the articles and reports by domain experts. The complied list of URLs from the previous step was studied by all authors for relevance. Most URLs were retained, with only a few exceptions that had no recognizable author and too little content to consider for analysis. The evaluation and interpretations are described in the following section.

5. Results and Discussion.

The growing number of legal regulation in all areas of social life has led also to an increasing importance and awareness of *compliance*, not only in the public sector but also – and even more intensively – in the business world.

One obvious sign is the new establishment of the position of the *Compliance Manager* in companies, as a result of the introduction of a *Compliance Management System (CMS)* by the international standard ISO 19600.

Compliance is generally understood as to act in accordance with the rules, which are set up by law in a very general sense, ranging from international convention, constitution to statutory law, but also including soft law, like technical standards (ISO, DIN, etc.)

or Corporate Governance Codes[6] and even ethical rules. This general definition shows already the complexity of the subject and it is therefore of utmost necessity for companies as well as stakeholders, to develop adequate algorithms which enable them to cope with the challenges which emerge with the increasing volume of compliance rules [20].

A practical and representative example for the importance of compliance is, that the biggest private bank in Germany, Deutsche Bank, recently made public, that it will increase its department *Compliance, Regulation and Combat against financial crime* by further 400 people, from now 2600 to 3000 at the end of the year 6. Many comments on that decision, voicing it would be strange, if so many people would supervise something, which is a matter of course (behaving according to the laws), obviously underestimate or ignore the complexity of the legal framework[7]. Everybody in the society, be it natural or legal person, is involved [18, 21, 22].

This complexity of rules, standards, regulations and policies, which as such are already quite complicated to comprehend on the national level, is in Europe overlaid by EU-law, generally summed up under the notion of *acquis communautaire*[8]. The *acquis* encompasses the whole body of European Union law applicable in the EU and is composed of actually more than 108,000, steadily increasing documents, including besides statutory law also EU Court verdicts and all kind of decisions taken by the various EU institutions. It especially includes the concept of primacy of EU law over national law and has therefore become a main target of the compliance management departments of private companies as well as in the public sector [23, 24, 25].

The recent EU legislation on data protection highlights the importance of compliance with EU law: The EU General Data Protection Regulation, which has been enforced on May 25, 2018, has brought the most important change in data privacy regulation in 20 years and was designed to harmonize data privacy laws to protect and empower all EU citizens' data privacy and to reshape the way organizations across the EU approach data privacy. In this context, the impact of the GDPR cannot be emphasized enough [26].

The biggest change to the regulatory landscape of data privacy has been introduced with the extended jurisdiction of the GDPR, as it applies to all companies, which are processing personal data of subjects residing in the Union, regardless of whether the processing takes place in the EU or not and no matter, if it is a small or multinational business. It is also not relevant, if the company's main business is data processing. For example, processing data in the human resources department of a company is also included, and business whose main operations are cloud-based will as well not be exempted from the GDPR enforcement. One additional criterion for application is, where the activities relate to: offering goods or services to EU citizens (irrespective of whether payment is required) and

[6]The German Corporate Governance Code (as amended on February 7, 2017) refers to compliance issues besides others as follows: 1. Foreword: The Code highlights the obligation of the Manadement and Supervisory Boards to ensure the continued existence of the company and its sustainable value creation in line with the principles of the social market economy (the company's best interest). These principles not only require compliance with the law, but also ethically sound and responsible behavior (the "reputable business person concept," Leitbild des ehrbaren Kaufmanns); 4.1.3 The Management Board ensures that all provisions of law and the company's internal policies are complied with, and endeavours to achieve their compliance by the group entities (Compliance). It shall also institute appropriate measures reflecting the company's risk situation (Compliance Management System, CMS). The increased demand on CMS, which companies are confronted with, has been described in a Deloitte Newsletter: `https://www.deloitte-tax-news.de/german-tax-legal-news/new-revision-of-the-german-corporate-governance-code-increased-demands-on-compliance-management-systems.html`. Last accessed October 16, 2018

[7]New item accessed June 2018 at `https://www.wr.de/wirtschaft/deutsche-bank-400-neue-mitarbeiter-fuer-compliance-abteilung-id214068305.html`. See also general statement and detailed compliance examples at `https://www.db.com/cr/en/concrete-compliance.htm`

[8]`https://ec.europa.eu/neighbourhood-enlargement/policy/glossary/terms/acquis_en`. Last accessed October 16, 2018

the monitoring of behavior that takes place within the EU. Non-EU businesses processing the data of EU citizens have to appoint a representative in the EU.

6. **Conclusions.** Compliance in the business world has been of interest for decades and gained new urgency with the implementation of the General Data Protection Regulation in the EU.

In this paper, the authors show the usage of big data mining of the social media platform Twitter in cooperation with legal and business domain expert evaluation.

Findings have shown that the information dissemination has not yet reached a wide understanding of the importance of the GDPR. Awareness and social responsibility are focus points of many organizations, but this is not reflected by social media perception.

Future Propositions. As the GDPR has just been implemented as of May 25, 2018, follow-up investigations are planned to track changes in perception.

Further text mining and a causality-based approach are necessary to minimize high-influence individuals skewing the perception.

REFERENCES

[1] T. W. Loe, L. Ferrell and P. Mansfield, A review of empirical studies assessing ethical decision making in business, *Journal of Business Ethics*, vol.25, pp.185-204, 2000.

[2] D. Harker, The importance of industry compliance in improving advertising self-regulatory processes, *Journal of Public Affairs*, vol.3, no.1, pp.63-75, 2002.

[3] D. E. Murphy, The federal sentencing guidelines for organizations: A decade of promoting compliance and ethics, *Iowa Law Review*, vol.87, pp.697-720, 2002.

[4] A. B. Carroll and K. M. Shabana, The business case for corporate social responsibility: A review of concepts, research and practice, *International Journal of Management Reviews*, vol.12, no.1, pp.85-105, 2010.

[5] V. L. Nielsen and C. Parker, Mixed motives: Economic, social, and normative motivations in business compliance, *Law and Policy*, vol.34, no.4, pp.429-462, 2012.

[6] S. Ghanavati, D. Amyot and L. Peyton, A systematic review of goal-oriented requirements management frameworks for business process compliance, *The 4th International Workshop on Requirements Engineering and Law (RELAW)*, pp.25-34, 2011.

[7] L. K. Treviño, G. R. Weaver and S. J. Reynolds, Behavioral ethics in organizations: A review, *Journal of Management*, vol.32, no.6, pp.951-990, 2016.

[8] D. Thornton, N. Gunningham and R. A. Kagan, Social license and environmental protection: Why businesses go beyond compliance, *Law Social Inquiry*, vol.29, no.2, pp.307-341, 2004.

[9] C. Hodges and R. Steinholtz, *Ethical Business Practice and Regulation: A Behavioural and Values-Based Approach to Compliance and Enforcement*, Hart Publishing, 2017.

[10] C. Hodges, *Law and Corporate Behaviour: Integrating Theories of Regulation, Enforcement, Compliance and Ethics*, Hart Publishing, 2015.

[11] N. Dvorak and W. E. Kruse, Managing employee risk requires a culture of compliance, *Gallup Business Journal*, 2016.

[12] W. S. Laufer, Social accountability and corporate greenwashing, *Journal of Business Ethics*, vol.43, pp.253-261, 2003.

[13] J. P. Albrecht, How the GDPR will change the world, *European Data Protection Law Review (EDPL)*, pp.287-289, 2018.

[14] C. Tankard, What the GDPR means for businesses, *Network Security*, vol.2016, no.6, pp.5-8, 2016.

[15] C. Tikkinen-Piri, A. Rohunen and J. Markkula, EU General Data Protection Regulation: Changes and implications for personal data collecting companies, *Computer Law & Security Review: The International Journal of Technology Law and Practice*, vol.34, no.1, pp.134-153, 2018.

[16] M. Cornock, General Data Protection Regulation (GDPR) and implications for research, *Maturitas*, vol.111, 2018.

[17] White Paper Supported by Gemalto, *Essential Security Technologies for GDPR Compliance*, Data Quality Management Group (DQM GRC), 2018.

[18] A. Calder, *EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide*, IT Governance Publishing, 2016.

[19] E. Grinschuk, *EU GDPR Compliance Compact: GDPR Checklist for Websites and Bloggers*, 2018.

[20] C. Basri, *Corporate Compliance*, Carolina Academic Press, 2017.

[21] S. P. Ramakrishna, *Enterprise Compliance Risk Management: An Essential Toolkit for Banks and Financial Services*, Wiley (John Wiley & Sons, Inc.), 2015.
[22] R. Walker, *Conflicts of Interest in Business and the Professions: Law and Compliance*, Thomson West, 2016.
[23] Directorate General for Competition (EU Commission), *Compliance Matters*, European Commission, 2013.
[24] M. Cremona, *Compliance and the Enforcement of EU Law*, Oxford University Press, 2012.
[25] A. Jakab and D. Kochenov, *The Enforcement of EU Law and Values: Ensuring Member States' Compliance*, Oxford University Press, 2017.
[26] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Springer, 2017.