Computer Supported Cooperative Work (CSCW) (2022) 31:33–77
https://doi.org/10.1007/s10606-021-09422-3

ECSCW CONTRIBUTION

Regional Differences in Information Privacy Concerns After the Facebook-Cambridge Analytica Data Scandal

Felipe González-Pizarro[1,3], Andrea Figueroa[2], Claudia López[1,*] & Cecilia Aragon[2]

[1]*Departamento de Informática, Universidad Técnica Federico Santa María, Santa Maria, Chile (E-mail: felipe.gonzalezp.12@sansano.usm.cl; E-mail: felipegp@cs.ubc.ca; E-mail: afigue@uw.edu; E-mail: claudia@inf.utfsm.cl; E-mail: aragon@uw.edu);* [2]*Human Centered Design and Engineering, University of Washington, Seattle, WA, USA;* [3]*Department of Computer Science, University of British Columbia, Vancouver, BC, Canada*

Accepted: 29 November 2021

Abstract. While there is increasing global attention to data privacy, most of their current theoretical understanding is based on research conducted in a few countries. Prior work argues that people's cultural backgrounds might shape their privacy concerns; thus, we could expect people from different world regions to conceptualize them in diverse ways. We collected and analyzed a large-scale dataset of tweets about the #CambridgeAnalytica scandal in Spanish and English to start exploring this hypothesis. We employed word embeddings and qualitative analysis to identify which information privacy concerns are present and characterize language and regional differences in emphasis on these concerns. Our results suggest that related concepts, such as regulations, can be added to current information privacy frameworks. We also observe a greater emphasis on data collection in English than in Spanish. Additionally, data from North America exhibits a narrower focus on awareness compared to other regions under study. Our results call for more diverse sources of data and nuanced analysis of data privacy concerns around the globe.

Key Words: Online privacy, Twitter, Word embedding, Content analysis, IUIPC

1. Introduction

The right to control one's personal information has gained significant importance lately (Lee et al. 2019). Indeed, 58% of the countries have data protection and privacy legislation, while another 10% have drafted legislation about it (United Nations Conference on Trade and Development 2020). This broad interest is related to the massive amount of personal data collected by information systems and the risk that such information could be wrongly distributed online (Lee et al. 2019).

The study of information privacy has advanced our understanding of individuals' concerns regarding organizational practices associated with collecting and using their personal information (Smith et al. 1996). However, a literature review revealed a strong bias towards USA-centered studies across privacy concerns literature and warned about the limitations to generalizability this entails (Bélanger and Crossler 2011; Okazaki et al. 2020). The review's authors hypothesized that individuals from different world regions have diverse cultures, values, and laws, which can, in turn, result in different conceptualizations of information privacy and its impacts (Bélanger and Crossler 2011; Mohammed and Tejay 2017). To study these differences, privacy research has often relied on survey-based studies (Cockcroft and Rekker 2016). For example, a questionnaire was applied to explore differences in privacy perceptions between Facebook users from Germany and the USA (Krasnova and Veltri 2010), and a cross-national survey was conducted to evaluate information attitudes of consumers in the USA and Brazil (Markos et al. 2017). These multi-country privacy studies have had limited sample sizes, which makes the results difficult to generalize (Lee et al. 2019; Huang and Bashir 2016). They also tend to be focused on one or two cultures, usually including the USA (Cockcroft and Rekker 2016). Hence, multi-country information privacy research is still needed to extend our understanding of this increasingly relevant topic around the globe (Adu et al. 2019; Zou et al. 2018).

We propose an alternative approach to study information privacy concerns over a large geographical scope. This work combines word embeddings, open coding, and content analysis to examine tweets related to a large data breach scandal. We seek to characterize similarities and differences in privacy terms across people who tweet about this issue in different languages and from different world regions. Inspired by Rho et al. (2018), where text analysis was used to analyze answers about individuals' privacy concerns, we analyze the semantic context in which privacy-related terms were used in tweets written by different groups of people.

We focus on the Facebook–Cambridge Analytica data scandal. In 2018, the firm Cambridge Analytica was accused of collecting and using the personal information of more than 87 million Facebook users without their authorization (Venturini and Rogers 2019; Isaak and Hanna 2018; Lapaire 2018). The scandal sparked multiple conversations over technology's societal impact and risks to citizens' privacy and well-being worldwide. Opinions, facts, and stories related to it took place on different social media platforms such as Twitter, where the hashtag #DeleteFacebook became a trending topic for several days (Lin 2018; Mirchandani 2018).

We analyze more than a million public tweets in Spanish or English that use hashtags or keywords related to the scandal. We divide the dataset by language (Spanish and English) and regions (Latin America, Europe, North America, and Asia) and create word-embeddings for each subset. Then, we systematically analyze and compare the semantic context of four keywords, such as *data*, *privacy*, *user*, and *company*, across the embeddings. We contrast our results with one of the most used information privacy concerns framework to find terms and tweets matching different concerns. Then, we test a null hypothesis that there is no difference in emphasis on information privacy terms across languages and world regions. In this process, we discover the presence of related concepts that could be integrated into information privacy frameworks, such as regulations. We also observe statistically significant language differences in emphasis on data collection and significant regional differences in emphasis on awareness. Finally, we discuss the implications of our results.

We summarize prior work on information privacy concerns in Section 2. Section 3 introduces our research question and hypothesis. Section 4 details our research method, while Section 5 reports on our findings. Section 6 offers a discussion of our results, limitations, and future work. Finally, Section 7 provides our conclusions.

2. Information Privacy Concerns

Information privacy concerns emerge when an individual “feels threatened by a perceived unfair loss of control over their privacy by an information-collecting body” (Lee et al. 2015). Previous research argues that information privacy concerns are a multidimensional construct (Jozani et al. 2020; Correia and Compeau 2017; Heravi and Mubarak 2018). A multidimensional approach allows identifying to what extent users are concerned about different aspects of information privacy (Yun et al. 2019; Hong and Thong 2013, 2013). Different authors have proposed alternative conceptualizations to measure information privacy concerns. We briefly review the most adopted ones in the following subsection. Then, we summarize prior work on differences in privacy concerns across countries, regions, and other characteristics.

Prior research has examined privacy concerns from other perspectives as well (Heravi and Mubarak 2018; Gerber et al. 2018; Yun et al. 2019). A vast portion of privacy research on social networking sites has focused on examining users’ privacy behaviors, such as the intention to provide personal information or transact online (Wisniewski et al. 2017; Kokolakis 2017; Gerber et al. 2018; Oghazi et al. 2020; Chai 2020; Heravi and Mubarak 2018; Markos et al. 2017). In a similar direction, several studies have investigated the use of privacy setting configurations (Wisniewski et al. 2015; Vitak et al. 2015; Wisniewski et al. 2017; Li et al. 2020). Rather than centering on behavior or behavioral intention, we focus our review on information privacy concerns that characterize general personal dispositions (Janice et al. 2013). We think that this part of the literature aligns better with what people can say, in a declarative way, about data privacy on social media.

2.1. Assessing Information Privacy Concerns

Two questionnaires have been widely used to evaluate individuals' information privacy concerns (Bélanger and Crossler 2011; Cockcroft and Rekker 2016; Morton and Angela Sasse 2014): Concerns for Information Privacy (CFIP) and Internet Users' Information Privacy Concerns (IUIPC).

2.1.1. CFIP: Concerns for Information Privacy

The CFIP framework (Smith et al. 1996) focuses on individuals' perceptions of how organizations use and protect personal information (Van Slyke et al. 2006). CFIP identifies four dimensions:

- *Collection:* concerns about personal data that is collected over time;
- *Unauthorized secondary use:* concerns about organizations using personal data for another purpose without the individual's authorization;
- *Improper access:* concerns about unauthorized people having access to personal data;
- *Errors:* concerns about adequate protections from deliberate and accidental errors in personal data.

To measure them, Smith et al. (1996) proposed and validated a 15-item questionnaire. The CFIP questionnaire was validated by surveying 355 consumers from the USA and applying confirmatory factor analysis (CFA) (Stewart and Segars 2002). So far, this questionnaire had been considered as one of the most established methods to measure quantitatively information privacy concerns (Harborth and Pape 2018) and had been widely used in the literature (Harborth and Pape 2017; Stewart and Segars 2002). However, the CFIP and its measurement instrument were originally defined for users in an offline context (Palos-Sanchez et al. 2017). As the Internet enabled new ways to collect and process data, it was expected that new concerns about information privacy might emerge (Malhotra et al. 2004), and a new framework was proposed: the IUIPC.

2.1.2. IUIPC: Internet Users' Information Privacy Concerns

Malhotra et al. (2004) introduced the IUIPC framework and conceptualized Internet users' concerns about information privacy from a perspective of fairness. Drawing from social contract theory (Janice et al. 2013), Malhotra et al. (2004) argue that personal data collection is perceived to be fair when a user has control over their personal data and is informed about the intentions that organizations have about how to use it. The IUIPC includes three constructs:

- *Collection:* concerns about the amount of personal data owned by others compared to the perceived benefits (Malhotra et al. 2004). It is related to the perceived fairness of the outcomes one receives.

  Users provide information if they expect to obtain something of value after a cost-benefit analysis of a transaction.

- *Control:* concerns about control over personal information, including approval, modification of collected data, and opportunity to opt-in or opt-out from data collection (Malhotra et al. 2004). It is related to the perceived fairness of the procedures that maintain personal data.
- *Awareness:* concerns about personal awareness of organizational information practices (Malhotra et al. 2004). It relates to issues of transparency of the procedures and specificity of information to be used.

A 10-item questionnaire to assess these constructs was validated in Malhotra et al. (2004). The questionnaire has been widely used to this day (Yun et al. 2019; Raber and Krüger 2018) because it considers the Internet context, and it can explain more variance in a person's willingness to transact than CFIP (Rowan and Dehlinger 2014). Recent work has explored text mining as an alternative research method to identify IUIPC dimensions. Raber and Krüger (2018) found that IUIPC dimensions can be derived from written text. They observed a correlation between IUIPC concerns, as measured by the questionnaire, and LIWC language features of social media posts from a sample of 100 users.

2.1.3. Other instruments of assessment

The Westin-Harris Privacy Segmentation Index measures individuals' attitudes and concerns about privacy and how they vary over time (Kumaraguru and Cranor 2005) based on answers to three questions (Egelman and Peer 2015a; Woodruff et al. 2014). It categorizes individuals into three groups (Kumaraguru and Cranor 2005; Da Veiga 2018; Motiwalla et al. 2014): *Fundamentalists* are highly concerned about sharing their data, protect their personal information, prefer privacy controls over consumer-service benefits, and are in favor of new privacy regulations; *Pragmatists* tend to seek a balance between the advantages and disadvantages of sharing personal information before arriving at a decision; *Unconcerned* users believe there is a greater benefit to be derived from sharing their personal information, trust organizations that collect their personal data and are the least protective of their privacy.

The Westin-Harris' index was introduced as a way to meaningfully classify internet users based on their attitude toward privacy and their motivations to disclose personal information (Torabi and Beznosov 2016). It has been used for several decades. However, recent studies have raised questions about its validity (Egelman and Peer 2015a). Prior work has failed to establish a significant correlation between the Westin-Harris' segmentation and context-specific, privacy-related actual or intended behaviors (Consolvo et al. 2005; Woodruff et al. 2014; Egelman and Peer 2015b).

The existence of a mismatch between privacy concerns and privacy behaviors, known as the "privacy paradox" (Kokolakis 2017; Dienlin and Trepte 2015), motivated the creation of a new measurement instrument. Buchanan's Privacy Concern scale aims to capture different aspects of the paradox. (Buchanan et al. 2007) developed three privacy scales: two of them assess privacy behavior, and the third one measures information privacy concerns. However, some limitations have been identified. Their scales are not able to identify different privacy dimensions, but only one, which appears to map onto the general concept of privacy concern. Thus, a more fine-grained examination is desirable to improve the design of this scale (Buchanan et al. 2007).

Because our study focuses on people's comments about a specific information privacy scandal (and not their privacy behavior), our work will mostly build upon the information privacy concerns frameworks, particularly the IUIPC.

2.2. Differences on Information Privacy Concerns

Information privacy concerns can vary across individuals based on peoples' perceptions and values (Buchanan et al. 2007). People may have different concerns even if they experience the same situation (Lee et al. 2015). It has been argued that information privacy concerns can be influenced by different factors (Smith et al. 2011), such as national culture (Cho et al. 2009; Huang and Bashir 2016; Cao and Everard 2008; Krasnova and Veltri 2010), and individuals' demographics (e.g., age, gender) (Zukowski and Brown 2007; Lee et al. 2019; Jai and King 2016; Rowan and Dehlinger 2014; Cho et al. 2009; Markos et al. 2017). We review these factors below.

2.2.1. National Culture

While there are similarities in what privacy means across cultures (Cockcroft and Rekker 2016), there is no universal consensus on its definition (Cannataci 2009). According to Newell (1995), several cultures do not possess an equivalent term to the English' privacy definition in their own language, e.g., Arabic, Dutch, Japanese, and Russian. Nevertheless, this does not mean that these cultures lack a sense of privacy (Newell 1995). Every society appreciates privacy in some way, but the expression of it varies (Cho et al. 2009).

The concept of national culture has been studied as one of the factors related to information privacy concerns (Nov and Wattal 2009; Malhotra et al. 2004; Bellman et al. 2004). National culture can be defined as "the collective mindset distinguishing the member of one nation from another" (Cho et al. 2009). Hofstede's cultural dimensions theory (Hofstede 1983) has been the most used conceptual model to study cultural differences in this context. This trend is expected since Hofstede's theory has been widely used to study the relationship between culture and technology (Leidner and Kayworth 2006), even though there are a number of criticisms of this theory (Terlutter et al. 2006). The latest version of this theory proposes six cultural dimensions (Hofstede 2011). Among them, the *individualism/collectivism* dimension has been found relevant to information privacy concerns. *Individualism/collectivism* refers to the extent to which individuals are part of groups beyond their immediate families.

Differences in information privacy concerns have been explained using some cultural dimensions at a country and regional level (see Table 1). Participants from individualistic countries (Australia and United States) exhibited a higher level of online privacy concerns than individuals from collectivist countries (Cho et al. 2009). The authors' rationale is that high individualism is associated with an emphasis on private life and independence from the collective; thus, people from individualist countries are more worried about privacy intrusions. In the same direction, Bellman et al. (2004) found that controlling for internet experience and privacy regulations, people from countries with high individualism show deeper concern about two CFIP dimensions: *unauthorized secondary use* and *improper access*.

On the other hand, no regional differences in privacy concerns were found through online surveys with 226 English-fluent crowd workers from six regions (Africa, Asia, Western Europe, Eastern Europe, North America, and Latin America). The authors argued that it is unclear if their finding is due to true similarities or a lack of enough power in measuring privacy concerns (Huang and Bashir 2016).

Table 1. Culture and information privacy concerns

Independent variables	Method	# participants and origin	Key findings
National culture	5-item questionnaire (Cho et al. 2009), based on a unidimensional conceptualization of online privacy concerns. Items were comprehensive enough to measure general concerns about online privacy.	1261 from Seoul, Singapore, Bangalore, Sydney, New York	Participants from individualistic countries exhibited higher concern about online privacy (Cho et al. 2009)
National culture	15-item questionnaire (CFIP) (Smith et al. 1996), based on a multidimensional constructive model of privacy concerns (collection, unauthorized secondary use, improper access, errors).	534 from 38 countries	Participants from individualistic countries showed higher concern about improper access and secondary use (Bellman et al. 2004)
Regional culture	4-item questionnaire (Dinev and Hart 2006), based on a unidimensional conceptualization of privacy concerns, which is defined as apprehension about how online personal information is used by others.	226 from Africa, Asia, Western and Eastern Europe, North and Latin America	No regional differences in privacy concerns were found (Huang and Bashir 2016)

2.2.2. Language

Relatedly, the Sapir-Whorf hypothesis suggests that the structure of anyone's native language influences the world-views they will acquire (Kay and Kempton 1984). Depending on the language, a message is coded and decoded differently based on standardized language norms and culture (Zarifis et al. 2019). Thus, individuals who speak different native languages could think, perceive reality and organize the world around them in different ways (Hussein 2012).

Previous work has explored how user-generated content can reveal different views about the same issues among people who write in different languages. Jiang et al. (2017) conducted a semantic network analysis to examine the semantic differences that emerge from the Wikipedia articles about China. Results suggest that Chinese-speaking and English-speaking contributors framed articles about China in different and even opposite ways, which were aligned to their national cultures and values. The Chinese version framed them from perspectives of authority respect, emphasizing harmony and patriotism. Articles in English were written from the point of view that is distinctive of many Western societies: the core value of democracy.

A potential role of the spoken language in the information privacy context has also been studied. Li et al. (2017) created a cross-cultural privacy prediction model. The model applies supervised machine learning to predict users' decisions on the collection of their personal data. Using answers from an online survey of 9,625 individuals from 8 countries on four continents: Canada, China, Germany, United States, United Kingdom, Sweden, Australia and, India,they found that the model's prediction accuracy improved when adding individual's language (English, Chinese, French, Swedish, and German) or Hofstede's cultural dimensions. Our work will build upon this line of reasoning to deepen our understanding of information privacy concerns across the globe.

2.2.3. Other individual characteristics

Even though our work will not address the relationship between demographics and information privacy concerns, we will briefly review the literature about this topic.

Prior studies suggest that older Internet users are more concerned about online information privacy than younger ones (Cho et al. 2009). Older participants were more sensitive to privacy issues and exhibited a greater desire to control the amount of information collected about them (Zukowski and Brown 2007). In contrast, younger users declared themselves to be more willing to share their personal information with third parties (Jai and King 2016).

The relation between privacy concerns and gender has also been studied (Cho et al. 2009). Jai and King (2016) found that women were less willing than men to permit third parties to share their personal information. Similarly, Rowan and Dehlinger (2014) observed that women reported greater information privacy concerns than their male counterparts. Both studies considered gender as binary.

Another relevant factor is participants' internet experience. As users grow in internet experience, concerns for online information privacy may decrease (Zukowski and Brown 2007). Bellman et al. (2004) concluded that participants with more internet experience were less concerned about online privacy overall, and in particular, were less worried about *improper access* and *secondary use*. This could be explained by increased familiarity with online privacy practices (Zukowski and Brown 2007).

3. Research Questions

Overall, while concepts around information privacy concerns have been extensively investigated, some limitations are shared among the studies that assess differences in these concerns worldwide. Most research has been conducted through surveys and has focused only on a few geographic regions, with a notable exception of Li et al. (2017). Many studies have had a limited sample size (Vitkauskaite 2010; Ur and Wang 2013; Ebert et al. 2020; Chai 2020; Li et al. 2020; Oghazi et al. 2020; Krasnova and Veltri 2010); thus, their findings' generalizability has been questioned (Lee et al. 2019). Moreover, when information privacy concerns questionnaires are delivered in English to speakers of other languages, key differences among countries may be obscured, as has happened with other cross-national research (Harzing and Maznevski 2002, 2006). Unfortunately, conducting larger-scale, multi-country, and multi-language surveys can be quite expensive (Harzing 2005; Li et al. 2020). Yet, large-scale research to deepen our understanding of information privacy concerns worldwide is still needed (Vitkauskaite 2010; Chai 2020; Li et al. 2020; Zou et al. 2018; Oghazi et al. 2020; Okazaki et al. 2020).

We seek to assess the feasibility of using social media data to identify information privacy concerns and characterize language and regional differences. Twitter is a popular micro-blogging service where individuals from different world regions who speak diverse languages share opinions, information, and experiences (Yaqub et al. 2017; Shen and Kuo 2014). Mining text from this platform has been used as a fast and inexpensive method to gather opinions from individuals (O'Connor et al. 2010), which can complement findings obtained from traditional polls or other research methods. Prior research has found a significant correlation between tweets and public opinion in diverse domains (O'Connor et al. 2010; Tumasjan et al. 2010; Ilyas et al. 2020; Tavoschi et al. 2020). Following this trend of research, we aim to investigate whether Twitter data can reveal people's information privacy concerns. Thus, our first research question is as follows:

- *RQ1: Which information privacy concerns are present over social media content about a data-breach scandal?*

As we have reviewed in the prior section, there are arguments and evidence to support that information privacy concerns can vary across culture, language, and demographics (Chai 2020; Li et al. 2020; Oghazi et al. 2020, 2020). If information privacy concerns are present in a Twitter dataset, we could explore how they differ across people who live in different parts of the world and those who speak different languages. As we do not expect any specific trend of differences, we propose to test the following null hypotheses:

– *H0a. There are no differences in information privacy concerns by language*
– *H0b. There are no differences in information privacy concerns by world region.*

4. Data & Methods

To answer our research question and test the hypotheses, we implemented a four-step methodology (see Fig. 1). We retrieved tweets associated with data privacy during a specific period (*Section 4.1. data collection*). We filtered the data, removing retweets and excluding tweets likely to be generated by bots (*Section 4.2. data pre-processing*). We created word-embeddings (a multi-dimensional representation of a corpus) for the remaining tweets according to their language and world region (*Section 4.3. text mining*). Finally, we conducted an analysis to identify similarities and differences in the semantic contexts of privacy keywords in the word embeddings (*Section 4.4. coding and analysis*). Details about each of these steps are presented below.

4.1. Data Collection

We retrieved tweets related to the Facebook and Cambridge Analytica scandal between April 1st and July 10th, 2018. We focused on tweets in Spanish and English.

On March 17, 2018, it was revealed that the data firm Cambridge Analytica used personal data of 87 million Facebook users for political advertising purposes without their consent (Schneble et al. 2018; Oghazi et al. 2020). This scandal caused the closure of Cambridge Analytica (Solon and Laughland 2018) and numerous lawsuits against Facebook in the USA and the European Union. On

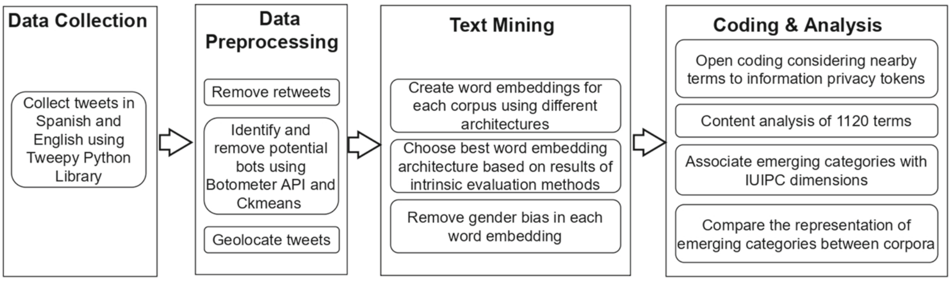

Data Collection
Collect tweets in Spanish and English using Tweepy Python Library
Data Preprocessing
Remove retweets
Identify and remove potential bots using Botometer API and Ckmeans
Geolocate tweets
Text Mining
Create word embeddings for each corpus using different architectures
Choose best word embedding architecture based on results of intrinsic evaluation methods
Remove gender bias in each word embedding
Coding & Analysis
Open coding considering nearby terms to information privacy tokens
Content analysis of 1120 terms
Associate emerging categories with IUIPC dimensions
Compare the representation of emerging categories between corpora

Figure 1. Methodology flow chart

Table 2. Datasets before and after data cleaning

Dataset	Spanish		English	
	#Tweets	#Accounts	#Tweets	#Accounts
Total	472,363	222,352	7,476,988	1,846,542
Original	106,656	47,951	1,572,371	574,452
With Botometer score	100,606	44,182	1,442,112	504,214
Human-owned	74,644	36,056	975,678	410,180

Twitter, a #DeleteFacebook campaign started as a response to this scandal (Lin 2018). As the Cambridge Analytica scandal triggered Twitter users from different world regions (who speak diverse languages) to spontaneously share their opinions, experiences, and perspectives about data privacy, we decided to use a sample of these tweets to answer our research question and test our hypotheses.

We used Tweepy[1] to collect relevant tweets. Tweepy is a Python library for accessing the standard real-time streaming Twitter API,[2] which allows to freely retrieve tweets that match a given query. If the query is too broad that it includes over 1% of the total number of tweets posted at that time worldwide, the query's response is sampled (Aghababaei and Makrehchi 2017; Morstatter et al. 2014). The way in which Twitter samples the data is unpublished. Nevertheless, studies have shown that as more data from the API is retrieved, a more representative sample of the Twitter stream is obtained (Leetaru 2019; Morstatter et al. 2013).

To obtain relevant tweets, we used Tweepy's language filter to retrieve tweets in Spanish or English. We manually crafted a list of hashtags and keywords related to the Cambridge Analytica scandal. We collected tweets that had at least one of these terms. Examples of these terms are: "#DeleteFacebook", "#CambridgeAnalytica", "#Mark Zuckerberg","Facebook", "Facebook Cambridge", and "Facebook data breach". Additionally, when appropriate, we added translations to Spanish of these terms to build the Spanish dataset.[3] In this way, if a tweet in Spanish had a hashtag in English, the tweet was collected and added to the Spanish dataset. A full list of the terms used to retrieve our data is available online[4].

Following this procedure, we retrieved more than 470,000 tweets in Spanish and more than 7.4 million tweets written in English (see Table 2). The tweets in Spanish were produced by approximately 220,000 users while tweets in English were generated by about 1.8 million unique Twitter accounts.

[1] http://www.tweepy.org/

[2] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters.html

[3] The authors are fairly confident of the quality of these translations because some of them are Spanish native speakers while others are English native speakers

[4] https://github.com/gonzalezf/Regional-Differences-on-Information-Privacy-Concerns

4.2. Data pre-processing

As we meant to analyze people's opinions about information privacy, we decided to pre-process our data in three ways. We removed all retweets to avoid analyzing exact duplicates. Afterwards, we sought to identify and filter out tweets that were generated by bots. Our last step was to associate tweets with different world regions. We further explain each of these steps below.

First, we excluded retweets to avoid analyzing exact duplicates of content. This methodology step is suggested by several authors (Hajjem and Latiri 2017; Agüero-Torales et al. 2021; Singh et al. 2020). We kept tweets, quoted tweets, and replies to tweets. Exclusion of retweets reduced our datasets' size by 80%. We refer to the resulting datasets as *original* tweets (see Table 2).

We used Botometer (Davis et al. 2016) to detect and remove tweets created by bots. Botometer uses machine-learning to analyse more than one thousand features (Badawy et al. 2018) including tweets' content and sentiment, accounts' and friends' metadata, retweet/mention network structure, and time series of activity (Varol et al. 2017; Yang et al. 2019) to generate a score that ranges from 0 to 1. A higher value suggests a high likelihood that an inspected account is a bot (Badawy et al. 2018). This tool has reached high accuracy (94%) in predicting both simple and sophisticated bots (Varol et al. 2017; Badawy et al. 2018). Botometer is free and has been widely used[5] (Varol et al. 2017; Yang et al. 2019).

Botometer processed all of the Twitter accounts who wrote original tweets. It returned a score for 44,182 (92.14%) and 504,214 (87.77%) accounts of the Spanish and English datasets, respectively. Botometer cannot generate scores for suspended accounts or those that have their tweets protected. We decided to remove the tweets from these accounts from our datasets because we cannot confidently claim that they come from humans' accounts. We applied the Ckmeans (Wang and Song 2011) algorithm to define a threshold to distinguish between humans' and bots' accounts. For each language, we clustered the Botometer scores into five groups, where the first group included the accounts with the lowest scores (more human-like) and the fifth group comprised those with the highest scores (more bot-like). After manually inspecting the accounts around the thresholds of each group, we concluded that the fourth and fifth groups in each dataset were unlikely to contain human accounts. Therefore, we used the fourth group's lowest threshold to discriminate humans' and bots' accounts. Accounts with a score lower than 0.4745 and 0.4947 in Spanish and English, respectively, were considered as human-owned. These thresholds are similar to those used in related work, where scores lower than 0.5 had been considered as humans (Varol et al. 2017; Badawy et al. 2018). As a result, our datasets contain 36,056 human-owned

[5] Since its release in May 2014, Botometer has served over one million requests (Davis et al. 2016) via its website (https://botometer.iuni.iu.edu) and its Python API (https://github.com/IUNetSci/botometer-python)

accounts that created 74,644 tweets in Spanish and 410,180 accounts that created 975,678 tweets in English.

Finally, we used the GeoNames API[6] to identify the country of residence of Twitter users in our datasets. On Twitter, users can self-report their city or country of precedence. Nevertheless, textual references to geographic locations can be ambiguous. For example, over 60 different places around the world are named "Paris" (Jackoway et al. 2011). To deal with this challenge, we employed the GeoNames API, which is a collaborative gazetteer project that contains more than 11 million entries and alternate names for locations around the world in a variety of different languages (Bergsma et al. 2013). Given a text, its algorithm performs operations to recognize potential locations, followed by a disambiguation process. This last step checks hierarchical relations and picks a location by their proximity to other locations mentioned in the text (Lagos 2017). This tool has yielded results with an accuracy above 80% (Jackoway et al. 2011).

We found that 80.68% of users in our Spanish dataset and 78.68% of users in our English dataset had filled the city or country fields in their profiles. However, the GeoNames API could not detect the users' location in several cases, for example when inaccurate information was provided (e.g., "Planet Earth.. where everyone else is from", "Mars"). Nonetheless, the tool was able to identify the location of users who created 58.7% of the Spanish tweets and 59.9% of the English ones. In the Spanish dataset, most tweets came from Spain (16.5%) and Latin American countries, such as Mexico (11.9%), Argentina (6.2%), and Venezuela (4.7%). In the English dataset, the majority of tweets came from the United States (32.4%), followed by United Kingdom (6.9%), India (3.2%), and Canada (2.7%) (see Table 3).

To compare information privacy concerns by geographical regions, we divided the Spanish Twitter dataset in two sets: tweets written by users from (1) Latin America and (2) Europe. Similarly, we categorized the English dataset into three groups: tweets written by users from (1) North America, (2) Europe and (3) Asia. As a result, five different language-regional datasets were generated. The Spanish-Latin America dataset includes 27,839 tweets written by 13,937 users. The Spanish-Europe dataset comprises 12,799 tweets created by 5,774 accounts. Regarding the English data, the North America dataset includes 342,400 tweets generated by 142,719 users, the English-Europe one has 111,745 tweets of 46,927 users, and the English-Asia dataset contains 42,208 tweets produced by 17,929 accounts (Table 4). We did not consider other subsets because of their small size. In Spanish, we only collected 1,929 tweets from North America and 217 tweets from Asia. In English, we only collected 3,851 tweets from Latin America.

[6] http://www.geonames.org/

Table 3. Top-10 most frequent user locations in the Spanish and English datasets

Spanish					English				
	Tweets		Users			Tweets		Users	
Location	#	%	#	%	Location	#	%	#	%
Spain	12,342	16.5	5,483	15.2	U.S	315,913	32.4	132,155	32.2
Mexico	8,852	11.9	4,720	13.1	U.K	66,901	6.9	29,656	7.2
Argentina	4,648	6.2	2,505	6.9	India	30,781	3.2	12,424	3.0
Venezuela	3,518	4.7	1,094	3.0	Canada	26,487	2.7	11,564	2.8
Colombia	2,447	3.3	1,348	3.7	Australia	13,375	1.4	6,501	1.6
U.S	1,823	2.4	948	2.6	Germany	9,260	0.9	3,493	0.9
Chile	1,806	2.4	1,073	3.0	France	8,605	0.9	3,006	0.7
Peru	1,116	1.5	619	1.7	Nigeria	5,156	0.5	2,787	0.7
Ecuador	893	1.2	455	1.3	U.A.E	5,120	0.5	1,504	0.4
Brazil	587	0.8	172	0.5	South Africa	4,962	0.5	2,912	0.7
Other	5,815	7.8	3,100	8.6	Other	98,100	10.1	42,658	10.4
Unknown	30,797	41.3	14,574	40.4	Unknown	391,018	40.1	161,767	39.4

Table 4. Tweets and users in each dataset

Language	Region	# of tweets	# of users
Spanish	Latin America	27,839	13,937
	Europe	12,799	5,774
English	North America	342,400	143,719
	Europe	111,745	46,927
	Asia	42,208	17,929

4.3. Text Mining: Word Embeddings to Identify Semantic Contexts

We employed word embeddings (Mikolov et al. 2013) to characterize the semantic context in which privacy-related keywords are framed. Based on co-occurrence of terms, word embeddings create a reduced multi-dimensional representation of a corpus. Such representation can be used to analyze the semantic proximity among the corpus' terms. Analyzing the closest terms of a given term can reveal the semantic context in which it is used Rho et al. (2018).

We created a set of word embeddings to enable cross-language and cross-regional comparisons. First, we built word embeddings for the Spanish and English datasets (containing both geolocated and non-geolocated tweets). Then, we generated word embeddings for each of our five language-regional datasets. Before creating the word embeddings, we transformed the text to lowercase. We also removed stop-words and digits from the tweets. We customized our stop-words to ensure that symbols like "#" were removed but not the words that follow it. Links and usernames were removed. Words with total frequency lower than three were ignored. These steps downsized the vocabulary by approximately 67% (details in Table 5).

We considered eight word embedding architecture combinations that involve *Word2Vec/FastText*, *CBOW/Skipgram* and different numbers of dimensions and epochs. As there is still no consensus about which word embedding evaluation

Table 5. Initial and final vocabulary size in each dataset

Language	Region	Initial vocabulary size	Final vocabulary size
Spanish	All	65,036	21,736
	Latin America	35,149	11,359
	Europe	21,630	6,696
English	All	244,371	76,128
	North America	115,710	41,109
	Europe	66,042	23,514
	Asia	39,120	13,896

method is more adequate (Wang et al. 2019), we evaluated each word embedding architecture for the English dataset over 18 intrinsic conscious evaluation methods (Wang et al. 2019) using a word embedding benchmark library.[7] (Wang et al. 2019) approach has categorized the evaluation methods in three categories:

– Word semantic similarity (WSS): RW, MEN, Mturk287, WS353R, WS353S, WS353, SimLex999, RG65 and TR9856
– Word Analogy (WA): Google Analogy Test set, MSR and SemEval 2012-2
– Concept categorization (CC): AP, BLESS, BM, ESSLI 1A, ESSLI 2B, and ESSLI 2C

To choose the best architecture, we designed a point system to reflect the embeddings' performance. For each evaluation method, the word embedding with the highest accuracy received a score of 8 points, the embedding with the second highest accuracy was assigned 7 points, and so on. After running all evaluation methods, we summed the points obtained for each architecture . Considering a negative sampling and windows size parameters equal to 5, a Word2Vec CBOW architecture with 300 dimensions trained during 50 epochs achieved the total highest score (see Table 6). The same architecture had the best performance for all English regional datasets. Given that these evaluation methods are not available for a Spanish corpus, the same architecture was used to create all the Spanish word embeddings.

Previous work has reported that word embeddings can reflect gender bias as a result of social constructs embedded in the data (Zhao et al. 2018; Jha and Mamidi 2017). To reduce gender bias while preserving its useful properties such as the ability to cluster related concepts, we followed (Bolukbasi et al. 2016) approach. This is a post-processing method that projects gender-neutral words to a subspace which is perpendicular to a gender dimension, defined by a set of terms associated with gender such as *girl*, *boy*, *mother* and *father* (Zhao et al. 2018). We applied the following procedure to our English embeddings: (1) we identified a gender subspace selecting pairs of English words that can reflect a gender direction in each word embedding such as *woman-man*, *daughter-son* and *female-male*, (2) we ensured that gender neutral words are zero in the gender subspace, and (3) we made neutral words equidistant to all pair of terms contained in a collection of equality sets. A equality set is composed by a pair of words that should differ only in the gender component such as {*grandmother, grandfather*} and {*guy, gal*}. During this process, we used the English terms suggested by Bolukbasi et al. (2016). For the Spanish word embeddings, we used Google Translate API[8] to translate the same terms.

[7] https://github.com/kudkudak/word-embeddings-benchmarks
[8] https://cloud.google.com/translate/

Table 6. Word embedding architectures and their evaluation scores

Architecture				Evaluation word embedding score			
Type	Model	Dim.	Epochs	WSS	WA	CC	Total
FastText	CBOW	100	10	22	19	25	66
Word2Vec	Skipgram	100	10	40	4	35	79
Word2Vec	CBOW	100	10	35	11	33	79
Word2Vec	CBOW	100	50	34	16	41	91
Word2Vec	CBOW	100	300	33	9	31	73
Word2Vec	CBOW	300	10	40	18	32	90
Word2Vec	**CBOW**	**300**	**50**	**53**	**21**	**36**	**110**
Word2Vec	CBOW	300	300	31	10	29	70

Best performance is indicated with bold font style

4.4. Manual coding & analysis

We conducted a systematic qualitative examination of the semantic contexts in which information privacy terms appear according to the word embeddings. First, we conducted open coding of the semantic neighborhoods of privacy-related keywords. After several iterations, we developed a set of categories to characterize them. To assess if information privacy concerns were present (RQ1), we contrasted these categories to a widely accepted framework to describe internet users' information privacy concerns.

We focused our investigation on four keywords in English: *information*, *privacy*, *users* and *company*. We used their corresponding translations in Spanish: *información*, *privacidad*, *usuarios* and *empresa*. We chose to include *information* and *privacy* because they are the main concepts under study. We could have added data; however, its semantic context is almost identical to that of information. Thus, adding it would have resulted in a mere duplication of terms. To increase the size of our dataset, we decided to add *users* and *company* because of their key roles in respect of controlling and safeguarding personal information. We also considered these terms more specific to the vocabulary of the data privacy domain than alternative ones (e.g., people, organizations).

For each embedding, we retrieved the closest terms to the four keywords. Closeness between each term and a keyword was measured using cosine similarity. For instance, the closest terms retrieved to the keyword *information* in the English word embedding were *info*, *data*, *details*, and *personal*, in that order. We chose to study the 40 closest terms after careful examination of the lists of close terms according to our different embeddings. After the 40th position in these lists, we rarely found terms that were even slightly related to information privacy. We reason that the value of this threshold is dataset-dependent. It is likely to be related to the vocabulary sizes (ours range from 6,696 to 41,109). In our case, we opted for using 40 as the threshold to study the semantic context of each keyword. Hence, we qualitatively analyzed 160 terms for each embedding. Overall, our dataset for qualitative analysis included 1,120 terms.

Two of the authors conducted open coding of the 320 terms retrieved from the Spanish and English word embeddings. Open coding is a process to identify, define and develop categories based on properties and dimensions of raw data (Williams and Moser 2019). We used this technique to identify distinct concepts and themes from the extracted terms (Williams and Moser 2019). After inspecting the retrieved terms during several iterations, the coders developed a coding guideline with multiple concept categories and their corresponding explanations to classify the retrieved terms. For example, the term *info* extracted from the keyword *information* was categorized as a *synonymous*, given that we can attribute to it the same meaning. The terms *data* and *details* were classified as *data & information*, and *personal* was labeled as *attribute or characteristic*. During a series of meetings, both coders compared their categorization process and refined a common coding guideline, establishing rules that would increase the categorization's reliability. The goal during this process is to segregate, group, regroup and re-link the terms to consolidate meaning and explanation of the categories (Williams and Moser 2019).

At the end of this process, 15 categories emerged from the data (see Table 10). Considering the four keywords, an inter-coder reliability measure (Cohen's kappa) of 0.685 and 0.754 were obtained for the Spanish and English dataset, respectively (see Table 7). These scores indicate substantial agreement (Viera et al. 2005) during the process.

We repeated the procedure for the regional datasets. The coders categorized the 40 closest terms to the keywords according to the coding guideline. Through an iterative process, a total of 800 words were manually coded. No new categories emerged from the data. On average, a Cohen's kappa above 0.722 was obtained in all the regional datasets (Table 7). These scores indicate substantial agreement (Viera et al. 2005) during the process.

To assess if information privacy concerns were present in a Twitter dataset about a data-breach scandal (RQ1), we compared the resulting categories with the IUIPC's dimensions: *collection*, *control*, and *awareness*. IUIPC is a theory-based model that has been widely used to study information privacy concerns on the internet (see Section 2.1.2). Then, we tested the null hypotheses about differences in information privacy concerns across language and world regions (H0a and H0b). To do so, we used a Chi-squared test to assess if the proportion of terms in the semantic contexts were significantly different across word embeddings. In all of these tests, we accounted for multiple comparisons by applying alpha adjustment according to Šidák (1967) and Haynes (2013). This method allowed us to control the probability of making false discoveries when performing multiple hypotheses tests.

Table 7. Inter-rater reliability (Cohen's kappa) score by dataset

Language	Region	Information	Privacy	Company	Users	Avg. by dataset
Spanish	All	0.864	0.630	0.604	0.642	**0.685**
	Latin America	0.749	0.673	0.687	0.778	**0.722**
	Europe	0.820	0.710	0.774	0.827	**0.783**
English	All	0.768	0.747	0.672	0.829	**0.754**
	North America	0.831	0.721	0.912	0.971	**0.859**
	Europe	0.777	0.743	0.805	0.807	**0.783**
	Asia	0.832	0.685	0.736	0.833	**0.771**
Average of all embeddings		**0.806**	**0.701**	**0.741**	**0.812**	**0.765**

Table 8. Top 20 closest terms to *information* and *privacy* in the Spanish and English word embeddings

Information		Privacy	
Spanish	English	Spanish	English
data	info	intimacy	data privacy
info	data	data	gdpr
third parties	details	confidentiality	protection
fact	personal	scams	users
third	users	personal data	user
interviewer	profiles	privacy policy	consumers
facebook	identifiers	digital security	data
users	personal data	data protection	transparency
privacy	private	identity	facebook
consent	records	minor	personal
authorization	user	facebook	consent
purposes	consent	third parties	security
private	advertisers	information	sharing
personal	permission	cibersecurity	data protection
location	metadata	sensitive	tos
ecomlancer	datas	emails	consumer
serve	companies	cookies	collection
viatec	individuals	protect yourself	opt
intimate	freely	suppose	trust
profiles	informations	take care of your data	privacyrights

5. Results

In this section, we address our research question and test the null hypotheses about differences in information privacy concerns by language and world regions.

As explained above, we create word embeddings for our Spanish and English datasets of tweets about the Cambridge Analytica scandal. Then, we take a closer examination of how the semantic context of four keywords varies across language and world regions. The semantic context is operationalized as the 40 closest terms of each keyword: *information*, *privacy*, *company*, and *users*. As an example, Table 8 and Table 9 show the 20 closest terms to the keywords, according to the Spanish and English word embeddings.[9] Full results are available online.[10]

[9] Terms in Spanish were translated to English by the authors

[10] https://github.com/gonzalezf/Regional-Differences-on-Information-Privacy-Concerns

Table 9. Top 20 closest terms to '*company* and *users* in the Spanish and English word embeddings

Company		Users	
Spanish	English	Spanish	English
company	companies	third parties	user
consultant	firm	sensitive	consumers
firm	companys	citizens	personal
organization	platform	illegally	peoples
obtain	firms	users	subscribers
relation	organization	authorization	customers
researcher	data	profiles	people
deliver	entity	private	facebook
plot	giant	used	data
finance	user	people	fb
ltd	facebook	illegal	apps
way	corporation	clients	individuals
facebook	fb	user	advertisers
illegally	organisation	improperly	privacy
companies	business	obtained	information
ca	users	nametests	app
own	businesses	information	citizens
brand	service	facebook	profiles
decide	ca	data	companies
scl	site	voters	collected
creole	personal	cambridgeanalytics	private
relations	organizations	infringement	consent
cambridge	employees	use	accounts
data	agency	purposes	permissions
laboratories	co	authorized	use

5.1. Information privacy concerns present in a Twitter dataset

As a result of the coding process, we define 15 categories to analyze the closest terms (see Table 10). To answer our first research question , we compare our categories with IUIPC, a framework widely used to measure information privacy concerns in the context of the Internet (Liu and Carter 2018). We find relationships among some of our categories and the three IUIPC concepts as well as our initial keywords, as shown in Fig. 2.

Table 10. Coding guideline to classify the semantic contexts of our keywords

Categories	Description	Spanish examples	English examples
Data & Information	Direct references to these concepts and examples of user data and its meaning	records, location, data, emails, profiles, contacts	profile, messages, documents, location, accounts, metadata
Companies	Entities that manipulate user data for their own purposes	facebook, cambridgeanalytica, scl, grindr, advertisers	facebook, cambridgeanalytica, google, emerdata, apple
Users	Data owners	users, consumers, population, citizens, people	users, consumers, subscribers, citizens, people
Data collection, handling and/or storage	Mechanisms and verbs associated with obtaining, collecting and handling data	use, obtained, log in, collecting, apps, mechanisms	collected, store, access, databases, usage, analyzed
Ownership agency	User-control over personal information	authorization, agree, consent, protect	consent, opt, shield, autonomy
Privacy & security terms	Words associated with data privacy and security	cybersecurity, confidentiality, intimacy, safe, secure	confidenciality, transparency, privately, dataprivacy, security
Security mechanisms	Tools and techniques that implement security services	credentials, password, biometry, key	encrypted, password, biometric, privacybydesign
Privacy & security risks	Entities or bad practices that can compromise sensitive data	trojan, cybercriminal, illegally, scams, stealing	grooming, databreach, misuse, illegally, violated
Regulation	Law, rule or regulation that controls the use of user data	rgpd, right to be forgotten, privacypolicy, iso, habeasdata	gdpr, tos, hippa, privacyrights, regulation
Synonymous	Same meaning than the keyword	info, company, firm, user, private	info, companies, firm, privacy, user,
Attribute or characteristic	A characteristic of the keyword	false, private, external, specialized, britain	sensitive, giant, holistic, strategic, affected

Table 10. (continued)

Categories	Description	Spanish examples	English examples
Action	Action or activity linked to the keyword	define, explode, promote, attend, move	order, solve, reveal, update, confirming
Third party	Can not be categorized as User or Company. There is not sufficient contextual information to do so	rrhh, medicians, interviewer, ex employee, philippine	government, agency, indians, europeans, developers,
Reaction or attitude	Way of feeling or acting toward a entity	guilt, honest, overfall, suffers, handle	willingly, freely, tighter, restricting, forced
Undetermined	Relationship between keyword and term is unknown	approximately, v.i.p, ground, higher, depth	psychological, millions, new, group, image

Yellow background is for categories that match our initial keywords. Light blue background is for categories related to IUIPC. Gray background is for other categories

Three categories match our initial keywords (Table 10, yellow background): (1) **data & information** is associated with the *information* keyword, including direct references to this concept and examples of user data and its meaning (e.g.,"messages" and "metadata"), (2) **companies** include terms about organizations that use personal data for their own purposes such as "Facebook" and "Apple", and (3) **users** contain references to this keyword (e.g., "customers", "people").

Five categories are related to IUIPC (Table 10, light blue background). We identify a **data collection, handling and/or storage** category that contains words associated with technology or techniques useful to obtain, collect or handle data

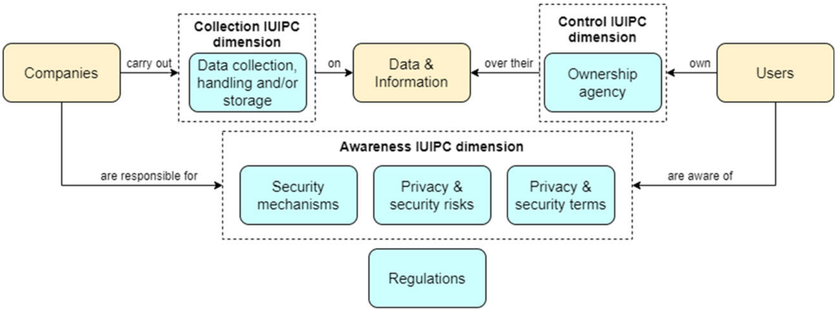


Figure 2. Relationships between our categories and IUIPC dimensions

(e.g: “databases”, “services”, “app”, “website”). This matches to the IUIPC’s *collection* dimension, which refers to the “degree to which a person is concerned about the amount of individual-specific data possessed by others relative to the value of benefits received’ ’(Malhotra et al. 2004). Examples of tweets that include terms that fit this category are:

> ‘@hidden_username @hidden_username This is bigger than facebook because all social media outlets collect and store this data on every user. If you haven’t looked, check and see what twitter has collected on you. Free apps are not free, neither are paid ones’

> ‘Facebook collects and sells PII data. Google and others maintain behavioral data anonymously and serve ads against it, but don’t connect that data to identities that are sold to advertisers. I was not aware Facebook was such an anomaly.’

The IUIPC’s *control* dimension denotes concerns about control over personal information. This is often exercised through approval, modification and opportunity to opt-in or opt-out (Malhotra et al. 2004). Terms related to this dimension appeared in the coding phase (e.g., “consent”, “opt”, “permission”) and were categorized as **ownership agency**. This category also includes advice directed to users and good privacy practices terms (e.g., “prevent”, “protect”). Examples of tweets related to this category are:

> ‘If anything we should learn from the #Facebook data breach. Don’t volunteer information and prevent that secondary data collection by using #adblocker and #VPN’

> ‘Cambridge Analytica whistleblower Christopher Wylie urges U.S. senators to focus less on data consent and more on the idea that it’s almost impossible to opt out of, for example, Google.’

The third IUIPC’s dimension is *awareness*, which refers to individual concerns about her/his awareness of organizational information privacy practices (Malhotra et al. 2004). Three of our categories are associated with this dimension: (1) **privacy and security terms** that include words associated with data privacy and security such as “confidentiality”, “transparency” and “safety”; (2) **security mechanisms** that refers to tools and techniques that implement security services (e.g., “password”, “encryption”); and, (3) **privacy & security risks** that denote entities or bad practices that can compromise sensitive data, for example: *“troyano”* (trojan), “databreach”, “grooming” and *“ciberdelincuente”* (cybercriminal). Tweets that use these terms are:

> ‘Hmm- what do you think? I forsee a wave of new social network startups- will any be able to rise? Besides privacy and transparency what else would you want from a social network? #swtech’

> ‘WhatsApp Co-Founder To Leave Company Amid Disagreements With Facebook. Facebook’s desire to weaken WhatsApp’s encryption and collect more personal data reportedly fueled the decision’

> ‘Canadian federal privacy officials warned that third-party developers’ access to Facebook users’ personal information raises serious privacy risks back in 2009. @hidden_link’

Another privacy-related category emerges from our coding but can not be easily associated with an IUIPC dimension. This is the **regulation** category, which includes terms associated with laws and rules that control the use of personal data such as “gdpr” in reference to the European General Data Protection Regulation or “tos” in reference to Terms of Services. Examples of tweets with these terms are:

> ‘New regulation in Europe called gdpr makes companies liable for data breaches with penalties which include fines of a percentage of global turnover. It feels like all Zuckerberg is liable for is a slap on the wrist and having to apologise in public’

> ‘#Today we are confirming that multiple snippets of data from CI that was lifted from facebook are in Russia. If you are an EU citizen this means you have a right to sue both companies for gdpr based infringements. We will be leading this cause should no one else step up....’

> ‘Senator to #Zuckerberg: Your terms of services are only a few pages long. People complain when online contracts are too long and filled with legalese. Now lawmakers are complaining they’re too short. What’s the threshold for length and detail, and how do we decide?’

Other categories are identified as well (Table 10, gray background). The **attribute or characteristic** category contains modifiers of a specific keyword. For example, the term “sensitive” emerges from the closest terms to *information*, and the term *“britanica”* (British) appears among the nearest terms to *company*. The **action** category includes words related to an act. For instance, the verbs *“obtener”* (obtain) and *“entregar”* (deliver) come out among the closest terms to *company*. The **third party** category contains terms related to entities that can not be categorized as user or company because there is not sufficient contextual information to do so, such as “indians”, “third”, “individuals” and “americans”. Additionally, the **reaction or attitude** category comprises terms that represent a way of feeling or acting toward a person, thing or situation. For example, the terms “deny” and “admitted” are present in the closest terms to *company*. A **synonymous** category emerge during the process as well. This contains equivalent terms to each keyword. For example, the terms “info” and “informations” are close to *information* and the terms “companys”, “corporation”, “companies” and “firm”

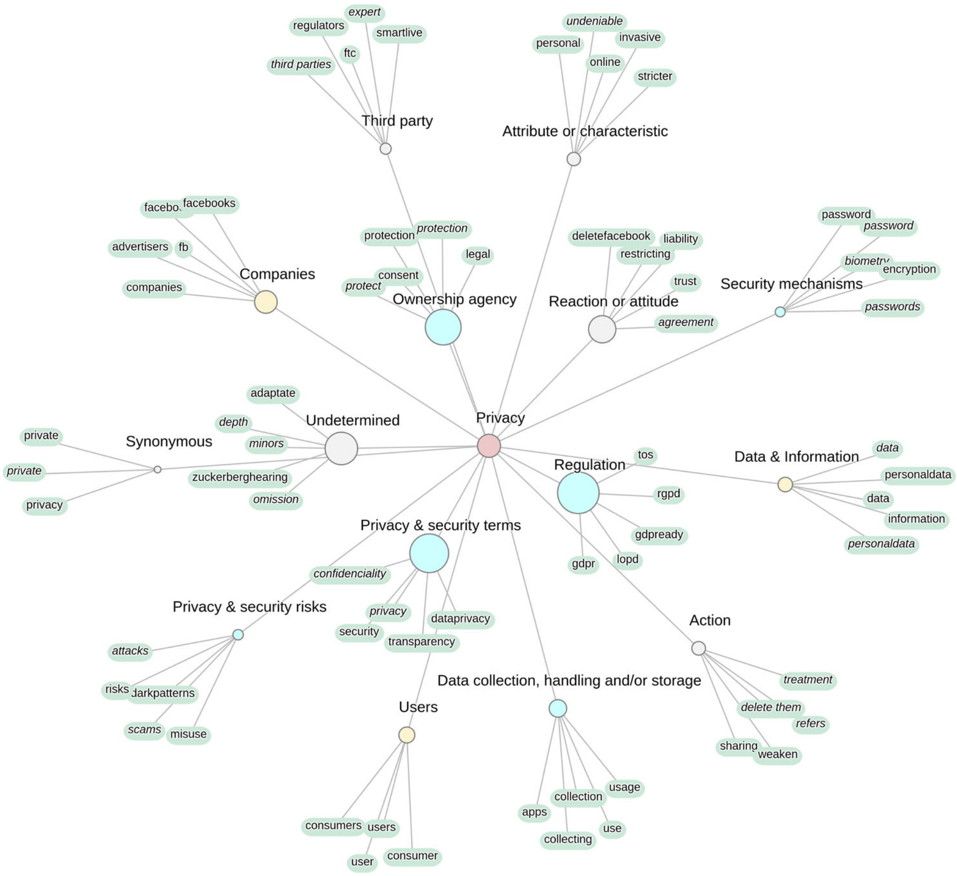
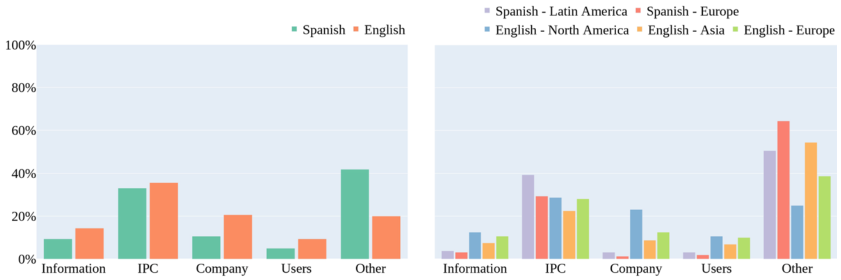


Figure 3. This force-directed graph represents the open coding categories related to the keyword *privacy* and provides examples of the terms that were coded as each category. Categories with the higher frequency are larger and closer to the keyword. Yellow nodes represent keywords and light-blue nodes denote privacy-related categories

appear among the closest terms to *company*. Terms with no clear relation to the keywords were classified as **undetermined**.

We used force-directed graphs (Kobourov 2013) to represent all the categories that emerged from the analysis of semantic context of each keyword. Figure 3 shows the categories related to the keyword *privacy*. In this graph, distance represents closeness in the semantic context. For example, terms that were categorized as regulation are closer to privacy than terms that were categorized as security mechanisms. Additionally, the visualization shows examples of terms in each category in Spanish or English.[11] Force-directed graphs of the categories and

[11] Terms in Spanish were translated to English by the authors. These terms are shown in cursive



(a) By Language

(b) By world-region

Figure 4. Proportion of categories by language (**a**) and world region (**b**)

terms associated with the other keywords (*information*, *company*, and *users*) are available online.[12]

Overall, we observe that the semantic contexts of four privacy-related keywords include terms corresponding to information privacy concerns. We illustrate such presence in Figure 2. We positioned each IUIPC dimension at the intersection between two of our keywords. *Companies* carry out collection, handling and or storage activities regarding *data* & *information*. *Users* exercise (some) agency over the control of their *data* & *information*. The awareness dimension arises from the *users*' perception of the *companies*' practices. Our results suggest that the awareness dimension might be further categorized into sub-topics, such as awareness of privacy and security terms, security mechanisms, and privacy and security risks.

Beyond what the IUIPC model proposes, we find that *regulations* are relevant to Twitter users who talk about Cambridge Analytica. We position this concept close to *awareness*, as it is considered an environmental factor that relates to information privacy concerns (Lee et al. 2019; Zou et al. 2018; Mohammed and Tejay 2017) but is not integrated into the IUIPC.

5.2. Emphasis on information privacy concerns across languages and world regions

As we are able to observe the presence of information privacy concerns on the Twitter datasets, we can now turn to test the null hypotheses regarding differences across language and world regions.

We compare the emphasis on information privacy concerns (IPC) in the semantic contexts that emerge from the different word embeddings. Figure 4 reports the distribution of terms that relate to the initial keywords (Table 10, yellow background) and IPC (Table 10, light blue background) in each language and world region under study. *Others* include all remaining categories.

[12] https://andreafigue.github.io/word_embeddings/visualization.html

To test our hypotheses, we performed Chi-square goodness-of-fit tests. Because we ran multiple tests, we applied Šidák correction to counteract the problem of multiple comparisons, thus controlling the family-wise error rate. According to our Šidák 's adjustment, to maintain an overall alpha of 0.05 for the collection of 10 tests, null hypotheses can be rejected when $p < 0.0102$.

We find no significant differences on the emphasis on information privacy concerns across languages or regions (see Table 11). Thus, we cannot reject the null hypotheses. We conclude that IPC are present at similar rates in Spanish and English. They cover a considerable proportion of the semantic contexts , with more than 30% of terms in both languages. Considering the regional datasets, IPC describes between 20% and 40% of the terms. While we observe some variation in emphasis on IPC across regions, with the largest proportion in the Latin American dataset and the smallest fraction in the Asian data, the differences across regions are not enough to be statistically significant.

The rest of the terms are better described by our initial categories, such as *company*, *information* and *users*. Compared to the IPC category, all of them cover smaller fractions of the semantic contexts under study. It should also be noted that irrelevant categories (grouped as *others*) add up to large proportions in all datasets, ranging from 30% to more than 60%. Together, these results reveal that while a social media dataset about a data breach scandal does bring relevant content about information privacy concerns, this comes with a fair amount of noisy content.

5.2.1. IUIPC dimensions

Digging deeper in the terms related to information privacy concerns, we analyze the proportions of terms that match each IUIPC dimension across languages and world regions (see Fig. 5 and Table 11).

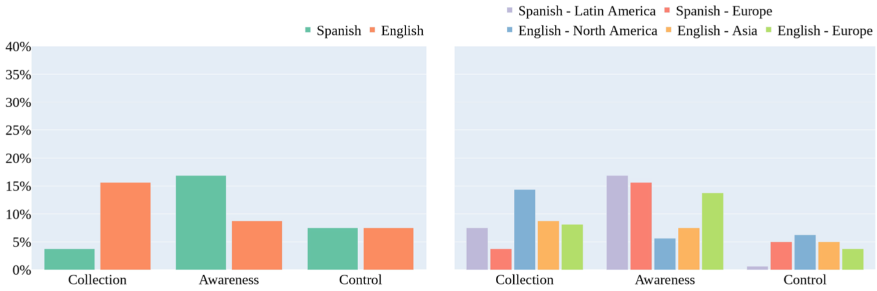
We observe a broader emphasis on *collection* in English ($\chi^2(1, 31) = 11.65, p =< .001$) than in Spanish. Cohen's effect size value (w = .61) suggests that this is a high practical significance (Cohen 1988). Even though this pattern seems to be influenced by a higher proportion on *collection* in the English content from North America than in any other region (Fig. 5), regional differences are not statistically significant after multiple comparisons correction ($\chi^2(4, 68) = 10.97, p = .03$).

In turn, while we cannot reject a null hypothesis regarding differences on *awareness* by language after corrections ($\chi^2(1, 41) = 4.12, p = .04$), we find a significant difference across world regions ($\chi^2(4, 95) = 13.58, p = .009$). Cohen's effect size value (w = .38) suggests a moderate to high practical significance. Here, we calculated the standard residuals to determine which world regions make the greater contribution to this chi-square test result. We find that compared with other world regions, data in English from North America have a smaller ratio of awareness terms (chi-square standard residual = −2.56). The

Table 11. Results of Chi-squared tests to compare proportions of terms by language and world regions

Null hypothesis	χ^2	N	DF	*p* value
There is no difference in % of *IPC* terms between languages	0.15	110	1	.70
There is no difference in % of *IPC* terms among world regions	8.04	237	4	.09
There is no difference in % of *collection* terms between languages	11.65	31	1	<**.001**
There is no difference in % of *collection* terms among world regions	10.97	68	4	.03
There is no difference in % of *control* terms between languages	0.00	24	1	1.00
There is no difference in % of *control* terms among world regions	7.15	33	4	.13
There is no difference in % of *awareness* terms between languages	4.12	41	1	.04
There is no difference in % of *awareness* terms among world regions	13.58	95	4	**.009**
There is no difference in % of *regulation* terms between languages	0.69	13	1	.41
There is no difference in % of *regulation* terms among world regions	11.69	26	4	.02

The null hypothesis was rejected if $p < .0102$



(a) By Language

(b) By world-region

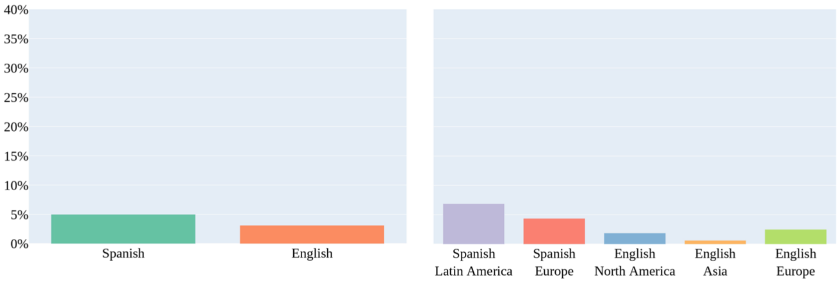
Figure 5. Proportion of terms related to the IUIPC dimensions by language (**a**) and region (**b**)

opposite is found in data in Spanish from Latin America (chi-square standard residual = 2.05).

Finally, we find no evidence to reject the null hypothesis regarding control. Control is equally present in both languages and the regions under study.

5.2.2. Regulation

Even though the concept of regulation is not part of the IUIPC dimensions, prior literature (Cockcroft and Rekker 2016; Lee et al. 2019; Da Veiga 2018) has suggested that it is related to people's concerns about information privacy. We find terms associated with this category in all our word embeddings (see Fig. 6). However, the difference in proportions between Spanish and English data is not statistically significant ($\chi^2(1, 13) = 0.69$, $p = .41$). Likewise, we do not find enough evidence to reject the null hypothesis regarding differences across world regions after multiple comparison correction ($\chi^2(4, 26) = 11.69$, $p = .02$) (see Table 11).



(a) By Language

(b) By world-region

Figure 6. Proportion of terms associated to regulations by language (**a**) and region (**b**)

6. Discussion

This work explores the potential of social media as a data source to study cross-language and cross-regional differences in information privacy concerns. We conduct an analysis of Twitter data related to a particular data breach news to deepen our understanding of how people from different world regions and who speak different languages frame privacy concerns. We chose to focus on the Cambridge Analytica scandal because it triggered a wide-ranging exchange on social media about user information and companies' data practices. We build upon the potential of word embeddings to derive a semantic context of each term in a corpus. The contexts are built according to terms that are commonly used in the same phrases. By characterizing a keyword's nearby terms, we seek to reveal the context in which a keyword was discussed (Rho et al. 2018). Based on more than a million non-duplicated, human-generated tweets, we generate word embeddings for data in Spanish and English and for data from Latin America, North America, Asia and Europe. For each embedding, we conduct a qualitative analysis of the semantic contexts of four privacy-related keywords: *information*, *privacy*, *company*, and *users*.

Collecting and analyzing the semantic contexts of these privacy-related keywords allows us to observe the presence of terms related to information privacy concerns in the collected tweets. Through iterative manual coding, we characterize the semantic contexts using 15 categories. Several of these categories are easily mapped to the three dimensions of the Internet User Information Privacy Concerns (IUIPC): *collection*, *awareness*, and *control* (See Fig. 2). In this way, we find evidence that social media content can reveal information about privacy concerns.

Our approach takes into consideration a vast amount of online content posted freely and spontaneously on Twitter to create the semantic context of each keyword. Thus, it gives a sense of a collective perspective on information privacy concerns by language and world region, which can become a complementary approach to current survey-based methods. Our method aims to discover knowledge from a large-scale social media dataset in a topic for which a ground truth does not exist. Unfortunately, such ground truth is unlikely to exist because large-scale, multi-country, and multi-language surveys are too expensive to conduct (Li et al. 2020). As an alternative approach, we used word embeddings to find the semantic contexts of relevant keywords and followed a qualitative approach to validate the results. We carefully analyzed more than a thousand terms of the semantic contexts, conducted open coding to formulate a data-grounded categorization, and then contrasted our categorization with IUIPC (Malhotra et al. 2004), one of the well-accepted theoretical conceptualizations of information privacy concerns. While this is not the common ground truth of other natural language processing tasks such as classification, our process draws from qualitative approaches to validate the results of an automated text analysis. We discuss below how our findings extend our current understanding of privacy concerns and open new lines of inquiry.

Beyond matching content to current conceptualizations of information privacy concerns, our results suggest a more granular categorization of one of them. Our results hint that *awareness* might include more specific sub-topics that users can be aware of, such as *privacy and security terms* (e.g., cybersecurity, confidentiality), *security mechanisms* (e.g., credentials, encrypted), and *privacy and security risks* (e.g., scams, grooming). The presence of terms that fit these categories reveals that they are already part of public online conversations around privacy. A distinction among broad privacy and security terms, mechanisms to protect data and potential data risks might be useful to further describe the kinds of knowledge people have. Additionally, awareness about some of these subtopics might be more influential than others. For example, knowing about risks and mechanisms might be a sign of higher privacy concerns while knowing broad privacy and security terms might not. The distinction between sub-topics could also guide users', educators' and practitioners' efforts to enhance information privacy literacy. Future work can explore the relevance of this distinction and its implications for information privacy practices.

Besides, the presence of the *regulation* category highlights its importance in relation to information privacy concerns. Regulation refers to laws or rules that aim to regulate the use of personal data. They have often been considered a factor influencing information privacy concerns(Ebert et al. 2020; Benamati et al. 2021). The emergence of this category from our open coding confirms its relevance through its frequent appearance in public posts about a data breach scandal. Such relevance might be related to the elaboration of laws and public policies about data usage worldwide. These regulations are not only a topic of data and law experts, but it seems to be part of the public discourse around data privacy online. It is noticeable the common presence of a specific regulation, the GDPR, in our datasets. GDPR is a privacy regulation that has been in effect since late May 2018 in the European Union (Gellman 2019). Our data collection period covered the early months of its implementation. This regulation prohibits processing and exploiting personal data such as health status, political orientation, sexual preferences, religious beliefs, and ethnic origin. Thus, it aims to decrease the privacy risks that may derive from malicious use of such information, including cases like the Cambridge Analytica scandal (Cabañas et al. 2018). GDPR seeks to convert individuals into empowered citizens involved in the decision-making process related to their personal information (Karampela et al. 2019). As an example, with this regulation in effect, companies are required to inform individuals about their rights (e.g., restriction of processing, erasure of data), the storage period of data, and additional sources that have been used to acquire personal data (Ebert et al. 2020). The explicit presence of GDPR in our data might be evidence of its influence on shaping people's arguments about privacy concerns and its importance not only in Europe but worldwide. Further work can focus on exploring how to integrate better the role of regulations into the current conceptualizations of information privacy concerns, which were proposed long before data privacy regulations were as common as they are now around the globe. Moreover, future work could also explore the interaction between regulations and specific information privacy concerns dimensions.

While we find similar rates of terms related to information privacy concerns across languages and regions, we observe significant differences in emphasis on collection and awareness. These results indicate that different groups view the Cambridge Analytica scandal from a particular standpoint. It is important to notice though that while information privacy terms appear through our method, they also come along with a considerable amount of other terms that we consider noisy data. Nevertheless, our findings show the potential of using social media data for cross-language and cross-regional comparisons to identify similarities and nuanced differences on privacy-related perspectives worldwide.

Our analysis reveals that the semantic contexts generated by tweets written in English have significantly more terms related to *collection* than those written in Spanish. This is a novel finding. When freely expressing online about privacy keywords, English speakers give significantly more emphasis to data collection than Spanish speakers. This difference can lead researchers and practitioners to explore the effectiveness of more tailored data privacy campaigns to specific populations. For example, populations that are more concerned about collection might need more information about the benefits of sharing their information to be able to make a decision about it. A high emphasis on collection in English is also congruent with prior literature observing that college students from the USA are more worried about collection of personal information than control over it (Yang 2013). Exploring if this trend is shared by people from other English speaking countries can help clarifying which of these patterns are better explained by location or language.

Future work can explore why we observe a significant language difference in emphasis on collection. A feasible explanation might be related to the users' country of residence. Note that our tweets in English come mainly from the USA and UK. Both were the countries most closely connected to the Cambridge Analytica scandal due to the misuse of data for political campaigns in the USA's 2016 presidential election and Brexit (Cadwalladr and Graham-Harrison 2018). It is possible that this shared experience resulted in a larger emphasis on collection in the English than the Spanish data. An alternative hypothesis is associated with differences in regulations. Information privacy concerns might be a reflection of customer privacy regulations in their respective countries (Markos et al. 2017; Kumar and Reinartz 2012). In contrast to European countries, that have adopted a data protection directive from a *government-imposed* perspective, the USA has followed an *industry self-regulation* (Kumar and Reinartz 2012). Considering that companies have more freedom to collect and process personal data in North America, it would be reasonable that data collection practices are of deeper concern to individuals from North America than those in Europe. This could also be supported by our data when observing that North America has the highest proportion of terms related to *company* (see Fig. 4), which we also found in our prior work using a different text mining method and a smaller dataset (González et al. 2019). However, our data analysis does not support the hypothesis of regional differences. It is possible that our data does not have enough power given the multiple comparisons we conducted. Future research is needed to explore alternative hypothesis that can explain the broader emphasis on collection among English speakers, compared to Spanish speakers.

We also observe significant regional differences on *awareness*. Particularly, data from North America shows the smallest emphasis on *awareness* while Latin America has the highest. Given that most studies on information privacy concerns are centered on the USA, this finding is particularly important. It warns us against the (sometimes implicit) assumption that North American data about privacy concerns can be generalizable to other regions. At least regarding emphasis on awareness, we find evidence that data from the USA is not similar to other regions. Thus, this result provides observational evidence to argue that it is necessary to include more diverse populations to have more a accurate understanding of the phenomena around data privacy. This finding also invites practitioners to address other regions, such as Latin America, using more different approaches in their terms of services and privacy policies. Populations that are more concerned about awareness might be more receptive to companies that use more transparent communications of their use of personal data, for example.

It is worth noting that Latin American shows the largest emphasis on *awareness*. Our results provide evidence of a disconnection between Latin America and North America regarding this aspect. It is possible that this broad interest on awareness can be a reflection of a connection of Latin America to the European perspective on data privacy. Latin America presents a particular scenario. It lies between two different approaches to personal data regulation: the principles contained by the European GDPR and the fragmented framework of the USA, where data protection is divided by sector (Aguerre 2019). Privacy regulations are considered an essential concern for many Latin American countries, and after data privacy breaches such as the Cambridge Analytica one, this issue has received increased attention in the public opinion and policy spheres in the region (Aguerre 2019). Previously, researchers have argued that GDPR could be one of the most influential pieces of data protection legislation ever enacted with influence beyond Europe (Kuner et al. 2017). Indeed, in Brazil, a new GDPR-like law (*Lei Geral de Proteção de Dados Pessoais, LGPD, in Portuguese*) has become effective since August 2020 (Dias Canedo et al. 2020). Future studies can explore connections among data privacy regulations worldwide, how they relate to public opinion on the issues of privacy, and how they are influenced by national and international data breaches.

As we found regional but not language differences in emphasis on privacy concerns, we conducted a follow-up analysis to assess whether there is a language difference within a single region. Europe was the only region where we had enough data in both languages to conduct such a comparison. We did not find significant differences in emphasis of any IUIPC dimension between data in English and Spanish from Europe ($\chi^2(3, 92) = 0.15$, $p = .98$). There was no evidence of significant differences in emphasis on regulations either. Thus, this additional analysis provides additional evidence to support that information privacy concerns are more related to the region of residence than the spoken language. Nevertheless, further research is required to understand better the role of regulatory regimes, consumer practices, and economic development factors on these differences (Okazaki et al. 2020). As the Spanish-English balance in tweets in our dataset is such that it does not lend itself to intra-region comparison for Asia and North and Latin America, future work could seek to explore if this pattern repeats in those regions as well

As with any study, our research has limitations. We collected data through the free standard streaming Twitter API using specific hashtags and keywords. Thus, we only had access to a limited sample of all the tweets about the scandal. We used Botometer to detect and remove tweets likely to be created by bots. This tool can only analyze Twitter public accounts; therefore, it could not be used on suspended accounts or those with their tweets protected when running our analysis. We decided to remove these accounts' tweets from our datasets because we can not confidently claim that humans generated them. Indeed, previous research suggests that it is likely that social bots were present in this cohort (Heredia et al. 2018). Moreover, we focused our investigation on four keywords in English: *information*, *privacy*, *users*, and *company* and their corresponding translations to Spanish. While using synonyms would have brought similar semantic contexts, adding more concepts can strengthen the results. Future work can explore other keywords such as: *intimacy*, and *consumers*. Similarly, we did not use the terms *user*, and *companies* as keywords. While word embeddings capture syntactic regularities such as singular/plural forms (Mikolov et al. 2013; Yeşiltaş and Güngör 2020), we reason that this methodological decision should not have affected considerably our results. Nevertheless, future work could include plural and singular versions of the same term to confirm this hypothesis. The sample size of our manual coding process (40 words per keyword in each embedding) could have impacted the results. We chose the number of retrieved terms after manually inspecting the list of nearest words by each keyword in all our embeddings. We picked a threshold that allowed us to obtain a high number of meaningful words in most embeddings. Higher thresholds make it more likely to include terms with no apparent relation to the keywords (e.g., v.i.p; ground; approximately). In word embeddings with reduced vocabularies like ours, the number of relevant terms available for a specific keyword is limited. This characteristic explains why the number of irrelevant terms (*Other* in Figure 3) is high in datasets with small vocabularies, such as the Spanish and English-Asia datasets. Future work could evaluate how sensitive our approach is to changes in vocabulary size and threshold for the nearest terms. This decision may introduce a bias in the results, and it is one of the limitations of our approach of social media textual data.

7. Conclusion

We conducted a cross-language and cross-regional study on social media content about a major data privacy leakage: the Cambridge Analytica scandal. We categorized our Twitter data into two different languages and four geographical regions.

Our results shed light on language and regional differences on information privacy concerns by 1) creating word embeddings by language and world regions to leverage social media data about a data breach scandal, 2) conducting open coding and content analysis of the semantic contexts (generated by the embeddings) of privacy-related keywords, 3) mapping the results to a well-known information privacy framework, and (4) conducting a comparative analysis across two languages and four world regions.

We found that data in English shows a broader emphasis on data collection , while data from North America shows the smallest emphasis on awareness. In turn, data from Latin America has the broadest emphasis on awareness. We discuss how our findings extend current conceptualizations of information privacy concerns, and how they might relate to regulations about personal data usage in the regions we analyzed.

Future work can dig deeper on the differences we observed and explore further the potential causes we discussed. Future studies might build upon our work to examine privacy concerns considering more languages, more geographical locations or different information privacy frameworks. Using our methodology to compare datasets across longer periods of time could be useful to determine if the semantic contexts of the privacy keywords changes over time.

Acknowledgments

The authors want to thank Francisco Tobar, MSc. Computer Science student at Universidad Técnica Federico Santa María, for helping us to strengthen our findings through statistical analysis. Moreover, we acknowledge anonymous reviewers for insightful comments that helped us revise and refine the paper.

Funding

This collaboration was possible thanks to the support of the Fulbright Program, under a 2017-18 Fulbright Fellowship award. This work was also partially funded by CONICYT Chile, under grant Conicyt/Fondecyt Iniciación/11161026. The first author acknowledges the support of the PIIC program from Universidad Técnica Federico Santa María and CONICYT-PFCHA/MagísterNacional/2019-22190332.

Compliance with Ethical Standards

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Adu, Ernest K; Nelly Todorova; and Annette Mills (2019). Do individuals in developing countries care about personal health information privacy? an empirical investigation. In *CONF-IRM 2019. Proceedings of the 2019 international conference on information resources management, auckland, new zealand, 27-29 may, 2019*. Auckland, New Zealand: School of Business. University of Auckland, p. 16.

Aghababaei, Somayyeh; and Masoud Makrehchi (2017). Activity-based Twitter sampling for content-based and user-centric prediction models. *Human-centric Computing and Information Sciences*, vol. 7, no. 1, p. 3.

Agüero-Torales, Marvin M.; David Vilares; and Antonio G. López-Herrera (2021). Discovering topics in Twitter about the COVID-19 outbreak in Spain. *Procesamiento del Lenguaje Natural*, vol. 66, pp. 177–190.

Aguerre, Carolina (2019). Digital trade in Latin America: mapping issues and approaches. *Digital Policy, Regulation and Governance*, vol. 21, no. 1, pp. 2–18.

Badawy, Adam; Emilio Ferrara; and Kristina Lerman (2018). Analyzing the digital traces of political manipulation: the 2016 Russian interference Twitter campaign. In *ASONAM 2018. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Barcelona, Spain, 28-31 August, 2018*. IEEE, pp. 258–265.

Bélanger, France; and Robert E. Crossler (2011). Privacy in the digital age: a review of information privacy research in information systems. *Management Information Systems Quarterly, MIS Quarterly*, vol. 35, no. 4, pp. 1017–1042.

Bellman, Steven; Eric J. Johnson; Stephen J. Kobrin; and Gerald L. Lohse (2004). International differences in information privacy concerns: a global survey of consumers. *The Information Society*, vol. 20, no. 5, pp. 313–324.

Benamati, John H.; Zafer D. Ozdemir; and H. Jeff Smith (2021). Information privacy, Cultural values, and Regulatory preferences. *Journal of Global Information Management (JGIM)*, vol. 29, no. 3, pp. 131–164.

Bergsma, Shane; Mark Dredze; Benjamin Van Durme; Theresa Wilson; and David Yarowsky (2013). Broadly improving user classification via communication-based name and location clustering on twitter. In *NAACL HLT 2013. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, 9-14 June, 2013*. Association for Computational Linguistics, pp. 1010–1019.

Bolukbasi, Tolga; Kai-Wei Chang; James Y. Zou; Venkatesh Saligrama; and Adam T. Kalai (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS 2016. Advances in Neural Information Processing Systems, Barcelona, Spain, 5-10 December, 2016*. NY. United States: Curran Associates Inc., pp. 4349–4357.

Buchanan, Tom; Carina Paine; Adam N Joinson; and Ulf-Dietrich Reips (2007). Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American Society for Information Science and Technology*, vol. 58, no. 2, pp. 157–165.

Janice, C. Sipior; Burke T. Wardm; and Regina Connolly (2013). Empirically assessing the continued applicability of the IUIPC construct. *Journal of Enterprise Information Management*, vol. 26, no. 6, pp. 661–678.

Cabañas, José González; Ángel Cuevas; and Rubén Cuevas (2018). Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. In *SEC'18. Proceedings of the 27th USENIX Conference on Security Symposium*. USENIX Association, pp 479–495.

Cadwalladr, Carole; and Emma Graham-Harrison (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. http://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Accessed 16 July 2020.

Cannataci, Joseph A. (2009). Privacy, technology law and religions across cultures. *Journal of information, Law and Technology, 2009*(1).

Cao, Jinwei; and Andrea Everard (2008). User attitude towards instant messaging: The effect of espoused national cultural values on awareness and privacy. *Journal of Global Information Technology Management*, vol. 11, no. 2, pp. 30–57.

Chai, Sangmi (2020). Does cultural difference matter on social media? an examination of the ethical culture and information privacy concerns. *Sustainability, 12*(19).

Cho, Hichang; Milagros Rivera-Sánchez; and Sun Sun Lim (2009). A multinational study on online privacy: global concerns and local responses. *New Media & Society*, vol. 11, no. 3, pp. 395–416.

Cockcroft, Sophie; and Saphira Rekker (2016). The relationship between culture and information privacy policy. *Electronic Markets*, vol. 26, no. 1, pp. 55–72.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences hillsdale*. L. Erlbaum Associates: N.J.

Consolvo, Sunny; Ian E. Smith; Tara Matthews; Anthony LaMarca; Jason Tabert; and Pauline Powledge (2005). Location disclosure to social relations: why, when, & what people want to share. In *CHI'05. Proceedings of the SIGCHI conference on Human factors in computing systems, Portland, Oregon, USA, 2-7 April, 2005*. New York, NY. USA: Association for Computing Machinery, pp. 81–90.

Correia, John; and Deborah Compeau (2017). Information Privacy Awareness (IPA): A Review of the Use, Definition and Measurement of IPA. In *HICSS-50. Proceedings of the 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Big Island, HI, USA, 4-7 January, 2017*. ScholarSpace / AIS Electronic Library (AISeL).

Da Veiga, Adéle (2018). An information privacy culture instrument to measure consumer privacy expectations and confidence. *Information & Computer Security*, vol. 26, no. 3, pp. 338–364.

Davis, Clayton Allen; Onur Varol; Emilio Ferrara; Alessandro Flammini; and Filippo Menczer (2016). Botornot: A system to evaluate social bots. In *WWW '16 Companion. Proceedings of the 25th International Conference Companion on World Wide Web, montréal, québec, Canada, 11-15 April, 2016*. International World Wide Web Conferences Steering Committee, pp 273–274.

Dias Canedo, Edna; Angelica Toffano Seidel Calazans; Eloisa Toffano Seidel Masson Masson; Pedro Henrique Teixeira Costa; and Fernanda Lima (2020). Perceptions of ICT Practitioners Regarding Software Privacy. *Entropy*, vol. 22, no. 4, p. 429.

Dienlin, Tobias; and Sabine Trepte (2015). Is the privacy paradox a relic of the past? an in-depth analysis of privacy attitudes and privacy behaviors. *European Journal of Social Psychology*, vol. 45, no. 3, pp. 285–297.

Dinev, Tamara; and Paul Hart (2006). An extended privacy calculus model for e-commerce transactions. *Information systems research*, vol. 17, no. 1, pp. 61–80.

Ebert, Nico; Kurt Alexander Ackermann; and Peter Heinrich (2020). Does context in privacy communication really matter? — a survey on consumer concerns and preferences. In *CHI'20. Proceedings of the 2020 CHI conference on human factors in computing systems, honolulu, HI, USA, 25-30 April, 2020*. New York: Association for Computing Machinery, pp. 1–11.

Egelman, Serge; and Eyal Peer (2015a). Predicting privacy and security attitudes. *ACM Special Interest Group on Computers and Society. SIGCAS*, vol. 45, no. 1, pp. 22–28.

Egelman, Serge; and Eyal Peer (2015b). Scaling the security wall: Developing a security behavior intentions scale (sebis). In *CHI'15. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18-23 April, 2015*. New York: Association for Computing Machinery, pp. 2873–2882.

Gellman, Robert (2019). Fair Information Practices: A Basic History-Version 2.19. Social Science Research Network. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2415020.

Gerber, Nina; Paul Gerber; and Melanie Volkamer (2018). Explaining the privacy paradox: a systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, vol. 77, pp. 226–261.

González, Felipe; Yihan Yu; Andrea Figueroa; Claudia López; and Cecilia Aragon (2019). Global reactions to the cambridge analytica scandal: a Cross-Language social media study. In *WWW '19. Companion proceedings of the 2019 world wide web conference, san francisco, USA, 13-17 May, 2019*. New York: Association for Computing Machinery, pp. 799–806.

Hajjem, Malek; and Chiraz Latiri (2017). Combining IR and LDA topic modeling for filtering microblogs. *Procedia Computer Science*, vol. 112, pp. 761–770.

Harborth, David; and Sebastian Pape (2017). Privacy concerns and behavior of pokémon go players in Germany. In *12th IFIP Summer School on Privacy and Identity Management, Ispra, Italy, 3-8 September, 2017.*, Springer, pp. 314–329.

Harborth, David; and Sebastian Pape (2018). German Translation of the Concerns for Information Privacy (CFIP) Construct. Social Science Research Network. *SSRN Electronic Journal*. https://ssrn.com/abstract=3112207.

Harzing, Anne-Wil (2005). Does the use of English-language questionnaires in cross-national research obscure national differences? *International Journal of Cross Cultural Management*, vol. 5, no. 2, pp. 213–224.

Harzing, Anne-Wil (2006). Response styles in cross-national survey research: a 26-country study. *International Journal of Cross Cultural Management*, vol. 6, no. 2, pp. 243–266.

Harzing, Anne-Wil; and Martha Maznevski (2002). The interaction between language and culture: a test of the cultural accommodation hypothesis in seven countries. *Language and Intercultural Communication*, vol. 2, no. 2, pp. 120–139.

Haynes, Winston (2013). *Bonferroni correction*. New York: Springer, pp. 154–154.

Heravi, Alireza; and Sameera Mubarak (2018). Kim-kwang Raymond Choo Information privacy in online social networks: Uses and gratification perspective. *Computers in Human Behavior*, vol. 84, pp. 441–459.

Heredia, Brian; Joseph D. Prusa; and Taghi M. Khoshgoftaar (2018). The Impact of Malicious Accounts on Political Tweet Sentiment. In *CIC'18. Proceedings of the 4th International Conference on Collaboration and Internet Computing, Philadelphia, PA, USA, 18-20 October, 2018*. Philadelphia: IEEE, pp. 197–202.

Hofstede, Geert (1983). National cultures in four dimensions: a research-based theory of cultural differences among nations. *International Studies of Management & Organization*, vol. 13, no. 1-2, pp. 46–74.

Hofstede, Geert (2011). Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture, 2*, article 8, 26.

Hong, Weiyin; and James Y. L. Thong (2013). Internet privacy concerns: an integrated conceptualization and four empirical studies. *Management Information Systems Quarterly. MIS Quarterly*, vol. 37, no. 1, pp. 275–298.

Huang, Hsiao-Ying; and Masooda Bashir (2016). Privacy by region: Evaluation online users' privacy perceptions by geographical region. In *FTC 2016. Future Technologies Conference, San Francisco, California, USA, 6-7 December, 2016*. IEEE, pp. 968–977.

Hussein, Basel Al-Sheikh (2012). The sapir-whorf hypothesis today. *Theory and Practice in Language Studies*, vol. 2, no. 3, pp. 642–646.

Šidák, Zbyněk (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 626–633.

Ilyas, Sardar Haider Waseem; Zainab Tariq Soomro; Ahmed Anwar; Hamza Shahzad; and Ussama Yaqub (2020). Analyzing Brexit's Impact Using Sentiment Analysis and Topic Modeling on Twitter Discussion. In DG.o'20. Proceedings of the 21st Annual International Conference on Digital Government Research, Seoul, Republic of Korea, 17-19 June, 2020. New York,: Association for Computing Machinery, pp. 1–6.

Isaak, Jim; and Mina J. Hanna (2018). User data privacy: facebook, Cambridge Analytica, and Privacy Protection. *Computer*, vol. 51, no. 8, pp. 56–59.

Jackoway, Alan; Hanan Samet; and Jagan Sankaranarayanan (2011). Identification of live news events using Twitter. In *LBSN '11. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, Illinois, USA, 1 November, 2011*. New York: Association for Computing Machinery, pp. 25–32.

Jai, Tun-Min Catherine; and Nancy J. King (2016). Privacy versus reward: Do loyalty programs increase consumers' willingness to share personal information with third-party advertisers and data brokers? *Journal of Retailing and Consumer Services*, vol. 28, pp. 296–303.

Jha, Akshita; and Radhika Mamidi (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science, Vancouver, Canada, August, 2017*. Association for Computational Linguistics, pp. 7–16.

Jiang, Ke; Grace A. Benefield; Junfei Yang; and George Barnett (2017). Mapping Articles on China in Wikipedia An Inter-Language Semantic Network Analysis. In *HICSS-50. Proceedings of the 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Big Island,, 4-7 January, 2017*. ScholarSpace / AIS Electronic Library (AISeL), pp. 2233–2242.

Jozani, Mohsen; Emmanuel Ayaburi; Ko Myung; and Kim-Kwang Raymond Choo (2020). Privacy concerns and benefits of engagement with social media-enabled apps: A privacy calculus perspective. *Computers in Human Behavior, 107*, article 106260, 15.

Karampela, Maria; Sofia Ouhbi; and Minna Isomursu (2019). Exploring users' willingness to share their health and personal data under the prism of the new GDPR: implications in healthcare. In *EMBC 2019. Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23-27 July, 2019*. IEEE, pp. 6509–6512.

Kay, Paul; and Willett Kempton (1984). What is the Sapir-Whorf hypothesis? *American anthropologist*, vol. 86, no. 1, pp. 65–79.

Kobourov, Stephen G. (2013). Force-Directed Drawing Algorithms. In: R. Tamassia (ed.) *Handbook of Graph Drawing and Visualization*: Chapman and Hall/CRC, pp. 383–408.

Kokolakis, Spyros (2017). Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. *Computers & security*, vol. 64, pp. 122–134.

Krasnova, Hanna; and Natasha F Veltri (2010). Privacy calculus on social networking sites: Explorative evidence from Germany and USA. In *HICSS'10. Proceedings of the 2010 43rd Hawaii*

International Conference on System Sciences, Koloa, Kauai, HI, USA, 5-8 January, 2010. Washington: IEEE, pp. 1–10.

Kumar, V.; and Werner Reinartz (2012). *Customer privacy concerns and privacy protective responses*. Berlin: Springer, pp. 279–300.

Kumaraguru, Ponnurangam; and Lorrie Faith Cranor (2005). *Privacy indexes: a survey of Westin's studies. Technical Report, CMU-ISRI-5-138*, Institute for Software Research International, School of Computer Science, Carnegie Mellon University, Pittsburgh. https://www.cs.cmu.edu/ponguru/CMU-ISRI-05-138.pdf. Accessed 16 July 2020.

Kuner, Christopher; Dan Jerker; B. Svantesson; Fred H. Cate; Orla Lynskey; Christopher Millard; and Nora Ni Loideain (2017). The GDPR as a chance to break down borders. *International Data Privacy Law*, vol. 7, no. 4, pp. 231–232.

Lagos, Vasileios (2017). *Comparative analysis of the hashtags #RefugeesWelcome and #StopRefugees: Tweets mining and examination with the Knime Analytics Platform*. Master's thesis, University of Peloponnese.

Lapaire, Jean-Rémi (2018). Why content matters. zuckerberg, Vox Media and the Cambridge Analytica data leak. *ANTARES: Letras e Humanidades*, vol. 10, no. 20, pp. 88–110.

Lee, Hwansoo; Dongwon Lim; Hyerin Kim; Hangjung Zo; and Andrew P Ciganek (2015). Compensation paradox: the influence of monetary rewards on user behaviour. *Behaviour & Information Technology*, vol. 34, no. 1, pp. 45–56.

Lee, Hwansoo; Siew Fan Wong; Jungjoo Oh; and Younghoon Chang (2019). Information privacy concerns and demographic characteristics: Data from a Korean media panel survey. *Government Information Quarterly*, vol. 36, no. 2, pp. 294–303.

Leetaru, Kalev (2019). Is Twitter's Spritzer Stream Really A Nearly Perfect 1% Sample Of Its Firehose?, Forbes. https://www.forbes.com/sites/kalevleetaru/2019/02/27/is-twitters-spritzer-stream-really-a-nearly-perfect-1-sample-of-its-firehose/. Accessed 16 July 2020.

Leidner, Dorothy E; and Timothy Kayworth (2006). A review of culture in information systems research: Toward a theory of information technology culture conflict. *Management Information Systems Quarterly. MIS Quarterly*, vol. 30, no. 2, pp. 357–399.

Li, Yao; Alfred Kobsa; Bart P Knijnenburg; and MH Carolyn Nguyen (2017). Cross-cultural privacy prediction. *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 2, pp. 113–132.

Li, Yao; Eugenia Ha Rim Rho; and Alfred Kobsa (2020). Cultural differences in the effects of contextual factors and privacy concerns on users' privacy decision on social networking sites. *Behaviour & Information Technology*, 1–23.

Lin, Yu-Wei (2018). #DeleteFacebook is still feeding the beast–but there are ways to overcome surveillance capitalism, The Conversation Trust. https://theconversation.com/deletefacebook-is-still-feeding-the-beast-but-there-are-ways-to-overcome-surveillance-capitalism-93874. Accessed 16 July 2020.

Liu, Dapeng; and Lemuria Carter (2018). Impact of citizens' privacy concerns on e-government adoption. In *DG-o'18. Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, 30 May - 1 June 2018*, pp. 1–6.

Malhotra, Naresh K.; Sung S. Kim; and James Agarwal (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, vol. 15, no. 4, pp. 336–355.

Markos, Ereni; George R. Milne; and James W Peltier (2017). Information sensitivity and willingness to provide continua: a comparative privacy study of the United States and Brazil. *Journal of Public Policy & Marketing*, vol. 36, no. 1, pp. 79–96.

Mikolov, Tomas; Ilya Sutskever; Kai Chen; Greg S Corrado; and Jeff Dean (2013). Distributed representations of words and phrases and their compositionality. In *NIPS 2013. Advances in neural information processing systems, Lake Tahoe, 5-10 December 2013*, pp. 3111–3119.

Mirchandani, Maya (2018). To delete, or not to #deleteFacebook, that is the question, The Wire. https://thewire.in/media/to-delete-or-not-to-deletefacebook-that-is-the-question. Accessed 16 July 2020.

Mohammed, Zareef A.; and Gurvirender P. Tejay (2017). Examining Privacy Concerns and Ecommerce Adoption in Developing Countries. *Computers & Security, 67*(C), 254–265.

Morstatter, Fred; Jürgen Pfeffer; and Huan Liu (2014). When is it biased?: assessing the representativeness of twitter's streaming API. In *WWW'14. Proceedings of the 23rd international conference on world wide web, Seoul, Korea, 7-11 April, 2014*. New York: Association for Computing Machinery, pp. 555–556.

Morstatter, Fred; Jürgen Pfeffer; Huan Liu; and Kathleen M. Carley (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM-2013. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, Massachusetts, USA, 8–11 July, 2013*. Menlo Park: AAAI Press, pp. 400–408.

Morton, Anthony; and M. Angela Sasse (2014). Desperately seeking assurances: Segmenting users by their information-seeking preferences. In *PST 2014. Proceedings of the 12th International Conference on Privacy, Security and Trust, Toronto, Canada, 23-24 July, 2014*. IEEE, pp. 102–111.

Motiwalla, Luvai F.; Xiaobai (Bob) Li; and Xiaoping Liu (2014). Privacy paradox: Does stated privacy concerns translate into the valuation of personal information? In *PACIS 2014. Proceedings of the 18th Pacific Asia Conference on Information Systems*, Chengdu, China, 24-28 June, 2014. p. 281 Article 281.

Newell, Patricia Brierley (1995). Perspectives on privacy. *Journal of Environmental Psychology*, vol. 15, no. 2, pp. 87–104.

Nov, Oded; and Sunil Wattal (2009). Social computing privacy concerns: antecedents and effects. In *CHI'09. Proceedings of the SIGCHI conference on human factors in computing systems, Boston, MA, USA, 4-9 April, 2009*. New York,: Association for Computing Machinery, pp. 333–336.

O'Connor, Brendan; Ramnath Balasubramanyan; Bryan R. Routledge; and Noah A. Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM-2010. Proceedings of the fourth international AAAI conference on weblogs and social media, Washington DC, USA, 23–26 May, 2010*. Menlo Park: AAAI Press, pp. 122–129.

Oghazi, Pejvak; Rakel Schultheiss; Koteshwar Chirumalla; Nicolas Philipp Kalmer; and Fakhreddin F. Rad (2020). User self-disclosure on social network sites: A cross-cultural study on Facebook's privacy concepts. *Journal of Business Research*, vol. 112, pp. 531–540.

Okazaki, Shintaro; Martin Eisend; Kirk Plangger; Ko de Ruyter; and Dhruv Grewal (2020). Understanding the Strategic Consequences of Customer Privacy Concerns: A Meta-Analytic Review. *Journal of Retailing*, vol. 96, no. 4, pp. 458–473.

Palos-Sanchez, Pedro; José Hernandez-Mogollon; and Ana Campon-Cerro (2017). The behavioral response to location based services: an examination of the influence of social and environmental benefits, and privacy. *Sustainability, 9*(11), article 1988.

Raber, Frederic; and Antonio Krüger (2018). Privacy Perceiver: Using Social Network Posts to Derive Users' Privacy Measures. In *UMAP'18. Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore, Singapore, 8-11 July, 2018*. New York: Association for Computing Machinery, pp. 227–232.

Rho, Eugenia Ha Rim; Gloria Mark; and Melissa Mazmanian (2018). Fostering Civil Discourse Online: Linguistic Behavior in Comments of #MeToo Articles across Political Perspectives. In CSCW 2018. Proceedings of the ACM on Human-Computer Interaction. *Association for Computing Machinery, 2*, article 147, 28.

Rowan, Mark; and Josh Dehlinger (2014). Observed gender differences in privacy concerns and behaviors of mobile device end users. *Procedia Computer Science*, vol. 37, pp. 340–347.

Schneble, Christophe Olivier; Bernice Simone Elger; and David Shaw (2018). The Cambridge Analytica affair and Internet-mediated research. *European Molecular Biology Organization. EMBO reports, 19*, no. 8, article e46579.

Shen, Chien-Wen; and Chin-Jin Kuo (2014). Analysis of social influence and information dissemination in social media: The case of Twitter. In *CISIM. Proceedings of the 13th IFIP International Conference on Computer Information Systems and Industrial Management, Ho Chi Minh City, Vietnam, November 2014*. Berlin: Springer Berlin Heidelberg, pp. 526–534.

Singh, Loitongbam Gyanendro; Akash Anil; and Sanasam Ranbir Singh (2020). SHE: Sentiment Hashtag embedding through multitask learning. *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 417–424.

Smith, H.Jeff; Tamara Dinev; and Heng Xu (2011). Information privacy research: an interdisciplinary review. *Management Information Systems Quarterly. MIS Quarterly*, vol. 35, no. 4, pp. 989–1016.

Smith, H Jeff; Sandra J. Milberg; and Sandra J. Burke (1996). Information privacy: measuring individuals' concerns about organizational practices. *Management Information Systems Quarterly, MIS Quarterly*, pp. 167–196.

Solon, Olivia; and Oliver Laughland (2018). Cambridge Analytica closing after Facebook data harvesting scandal, The Guardian. https://www.theguardian.com/uk-news/2018/may/02/cambridge-analytica-closing-down-after-facebook-row-reports-say. Accessed 16 July 2020.

Stewart, Kathy A.; and Albert H. Segars (2002). An empirical examination of the concern for information privacy instrument. *Information Systems Research*, vol. 13, no. 1, pp. 36–49.

Tavoschi, Lara; Filippo Quattrone; Eleonora D'Andrea; Pietro Ducange; Marco Vabanesi; Francesco Marcelloni; and Pier Luigi Lopalco (2020). Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines & Immunotherapeutics, 16*(5), 1062–1069. PMID 32118519.

Terlutter, Ralf; Sandra Diehl; and Barbara Mueller (2006). The GLOBE study—applicability of a new typology of cultural dimensions for cross-cultural marketing and advertising research. *International advertising and communication*, 419–438.

Torabi, Sadegh; and Konstantin Beznosov (2016). Sharing health information on facebook: practices, Preferences, and Risk Perceptions of North American Users. In *SOUPS 2016. Twelfth Symposium on Usable Privacy and Security. USENIX Association*, 301–320.

Tumasjan, Andranik; Timm O. Sprenger; Philipp G. Sandner; and Isabell M. Welpe (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM-2010. Proceedings of the Fourth international AAAI conference on weblogs and social media, Washington DC, USA, 23–26 May, 2010*. Menlo Park, California: AAAI Press.

United Nations Conference on Trade and Development (2020). Data Protection and Privacy Legislation Worldwide. https://unctad.org/en/Pages/DTL/STI_and_ICTs/ICT4D-Legislation/eCom-Data-Protection-Laws.aspx. Accessed 12 February 2020.

Ur, Blase; and Yang Wang (2013). A cross-cultural framework for protecting user privacy in online social media. In *WWW '13. Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13-17 May 2013*. New York: Association for Computing Machinery, pp. 755–762.

Van Slyke, Craig; J. T. Shim; Richard Johnson; and James J. Jiang (2006). Concern for information privacy and online consumer purchasing. *Journal of the Association for Information Systems*, vol. 7, no. 6, pp. 415–444.

Varol, Onur; Emilio Ferrara; Clayton A. Davis; Filippo Menczer; and Alessandro Flammini (2017). Online human-bot interactions: detection, estimation, and characterization. In *ICWSM-2017. Proceedings of the eleventh international AAAI conference on web and social media, Montréal, Québec, Canada, 15–18 May, 2017*. Palo Alto: AAAI Press, pp. 280–289.

Venturini, Tommaso; and Richard Rogers (2019). "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, pp. 1–9.

Viera, Anthony J.; Joanne M. Garrett et al. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, vol. 37, no. 5, pp. 360–363.

Vitak, Jessica; Stacy Blasiola; Sameer Patil; and Eden Litt (2015). Balancing audience and privacy tensions on social network sites: Strategies of highly engaged users. *International Journal of Communication*, vol. 9, pp. 1485–1504.

Vitkauskaite, Elena (2010). Overview of research on cross-cultural impact on social networking sites. *Economics and Management*, vol. 15, pp. 844–848.

Wang, Bin; Angela Wang; Fenxiao Chen; Yuncheng Wang; and C.-C. Jay Kuo (2019). Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing, 8*, article e19, 14.

Wang, Haizhou; and Mingzhou Song (2011). Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, vol. 3, no. 2, pp. 29–33.

Williams, Michael; and Tami Moser (2019). The art of coding and thematic exploration in qualitative research. *International Management Review*, vol. 15, no. 1, pp. 45–55.

Wisniewski, Pamela J.; A. K. M. Najmul Islam; Bart P. Knijnenburg; and Sameer Patil (2015). Give Social Network Users the Privacy They Want. In *CSCW'15. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Vancouver, BC, Canada, 14-18 March, 2015*. New York: Association for Computing Machinery, pp. 1427–1441.

Wisniewski, Pamela J.; Bart P. Knijnenburg; and Heather Richter Lipford (2017). Making privacy personal: Profiling social network users to inform privacy education and nudging. *International Journal of Human-Computer Studies*, vol. 98, pp. 95–108.

Woodruff, Aison; Vasyl Pihur; Sunny Consolvo; Laura Brandimarte; and Alessandro Acquisti (2014). Would a Privacy Fundamentalist Sell Their DNA for $1000... If Nothing Bad Happened as a Result? The Westin Categories, Behavioral Intentions, and Consequences. In *SOUPS 2014. Proceedings of the 10th Symposium On Usable Privacy and Security, Menlo Park, CA, USA, 9-11 July, 2014*. USENIX Association, pp. 1–18.

Yang, Hongwei (2013). Young American consumers' online privacy concerns, trust, risk, social media use, and regulatory support. *Journal of New Communications Research*, vol. 5, no. 1, pp. 1–30.

Yang, Kai-Cheng; Onur Varol; Clayton A. Davis; Emilio Ferrara; Alessandro Flammini; and Filippo Menczer (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, vol. 1, no. 1, pp. 48–61.

Yaqub, Uama; Soon Ae Chun; Vijayalakshmi Atluri; and Jaideep Vaidya (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. *Government Information Quarterly*, vol. 34, no. 4, pp. 613–626.

Yeşiltaş, Gökçe; and Tunga Güngör (2020). Intrinsic and Extrinsic Evaluation of Word Embedding Models. In *ASYU 2020. Proceedings of the 2020 Innovations in intelligent systems and applications conference, Istanbul, 15-17 October 2020*, pp. 1–6.

Yun, Haejung; Gwanhoo Lee; and Dan J. Kim (2019). A chronological review of empirical research on personal information privacy concerns: an analysis of contexts and research constructs. *Information & Management*, vol. 56, no. 4, pp. 570–601.

Zarifis, Alex; Richard Ingham; and Julia Kroenung (2019). Exploring the language of the sharing economy: Building trust and reducing privacy concern on Airbnb in German and English. *Cogent Business & Management, 6*(1), article 1666641.

Zhao, Jieyu; Yichao Zhou; Zeyu Li; Wei Wang; and Kai-Wei Chang (2018). Learning Gender-Neutral word embeddings. In *EMNLP 2018. Proceedings of the 2018 conference on empirical* *methods in natural language processing, brussels, belgium, 31 october – 4 november, 2018*. Association for Computational Linguistics, pp. 4847–4853.

Zou, Yixin; Abraham H. Mhaidli; Austin McCall; and Florian Schaub (2018). “I’ve Got nothing to lose”: Consumers’ risk perceptions and protective actions after the equifax data breach. In *SOUPS 2018. Fourteenth symposium on usable privacy and security, baltimore, MD, USA, 12–14 August, 2018*. Baltimore: USENIX, Association, pp. 197–216.

Zukowski, Tomasz; and Irwin Brown (2007). Examining the influence of demographic factors on internet users’ information privacy concerns. In *SAICSIT ’07. Proceedings of the 2007 Annual Conference of the South African Institute of Computer Scientists and Information Technologists, Port Elizabeth, South Africa, 2-3 October 2007*. New York: Association for Computing Machinery, pp. 197–204.

Publisher’s Note