Global Reactions to the Cambridge Analytica Scandal: An Inter-Language Social Media Study

Felipe González
UTFSM
Santiago, Chile
felipe.gonzalezpi@usm.cl

Yihan Yu
University of Washington
Seattle, United States
yyu2016@uw.edu

Andrea Figueroa
UTFSM
Valparaíso, Chile
andrea.figueroa@usm.cl

Claudia López
UTFSM
Valparaíso, Chile
claudia@inf.utfsm.cl

Cecilia Aragon
University of Washington
Seattle, United States
aragon@uw.edu

ABSTRACT

Currently, there is a limited understanding of how data privacy concerns vary across the world. The Cambridge Analytica scandal triggered a wide-ranging discussion on social media about user data collection and use practices. We conducted an inter-language study of this online conversation to compare how people speaking different languages react to data privacy breaches. We collected tweets about the scandal written in Spanish and English between April and July 2018. We used the Meaning Extraction Method in both datasets to identify their main topics. They reveal a similar emphasis on Zuckerberg's hearing in the US Congress and the scandal's impact on political issues. However, our analysis also shows that while English speakers tend to attribute responsibilities to companies, Spanish speakers are more likely to connect them to people. These findings show the potential of inter-language comparisons of social media data to deepen the understanding of cultural differences in data privacy perspectives.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Information systems** → *Document topic models.*

KEYWORDS

Data privacy; inter-language comparisons; Twitter

1 INTRODUCTION

While there is evidence that concerns about privacy and its intricate relationship with users' decisions to use social media have

been rising among Americans [30], much less is known about these perspectives across the world. So far, most of the literature about privacy concerns has been focused on the United States and even though recently research has begun to be carried out examining European privacy perspectives after the General Data Protection Regulation (GDPR) implementation, privacy research in international settings is still needed [29]. A few survey studies of people from different pairs of countries have been conducted to address. This, and the results so far indicate that data privacy perspectives vary significantly across countries. Belgians self-reported lower levels of concerns about sensitive information leakage than people from the United States. Harris et al. attributed this difference to the privacy laws of the respondents' countries of origin [13, 28]. A survey of North American and Turkish freshmen living in similar residence hall settings [21] showed that Americans wished for more privacy in their hall rooms than Turkish students. In the context of e-commerce, another survey found that Italians tend to exhibit lower privacy concerns than Americans [10].

Unfortunately, relying solely on survey data for cross-cultural studies of data privacy has various limitations. Most of them focus only on two geographic regions and have limited sample size [33]. Additionally, most privacy surveys are only available in English and only a few of them have been translated to other languages [33]. Surveys of a multinational or global nature that can mitigate these limitations would be very costly, which makes it difficult to compare privacy attitudes more broadly.

As the Cambridge Analytica scandal unfolded and people became aware that the personal data of 87 million Facebook users were exposed without their consent and used by Cambridge Analytica to support political campaigns [22], thousands of people in different parts of the world expressed on social media their reactions to and reflections on the scandal, its relationship to data privacy, and its broader implications. Indeed, a movement to delete Facebook accounts emerged and the #deleteFacebook hashtag was trending for several days [25].

In this paper, we report on a study that observes Twitter activity about the Cambridge Analytica scandal in Spanish and English and proposes a methodology for inter-language comparison of social media text. We believe that our approach offers an alternative or complementary method to conduct studies on data privacy perspectives across speakers of different languages and may provide a roadmap for future cross-cultural research. As Twitter allows people

to express themselves freely and spontaneously and in different languages, it enables a unique opportunity to analyze multi-language large-scale data. These characteristics allow researchers to address some of the limitations of the survey-based methods described above, such as those related to language and sample size [36]. Our premise is that written communication can be a "window into culture and an external reflection of cultural values" [7]; therefore, what people write about the scandal in their own languages can reveal differences and similarities in data privacy concerns across users worldwide.

We summarize prior work on inter-language comparisons of social media in Section 2. Section 3 introduces our research question, Section 4 details our research method, Section 5 reports on our findings, and Section 6 offers a discussion of our results, its limitations, and future work. Finally, Section 7 provides our conclusions.

2   INTER-LANGUAGE COMPARISONS OF USER-GENERATED WRITTEN CONTENT

Prior research has used social media text in different languages to make comparisons among people who speak these languages. An analysis of more than 62 million tweets compared the top 10 most common languages regarding the use of features such as: URLs, hashtags, mentions, replies and re-tweets [17]. The findings show that German-speaking users tend to include more URLs and hashtags in their tweets than other users, while Korean-speaking users are prone to reply to each other more often than speakers of other languages. Hong et al. argue that users of different languages use Twitter for different purposes. The German community often uses this platform for information sharing, while the Korean community employs it for conversational purposes. Another study analyzes tweets written by Americans in English and Japanese in their official language [1]. As opposed to Americans, Japanese people tweet more self-related messages and more messages about TV programs. In turn, Americans tweet more about their peers, sports and news.

Previous work has also explored how user-generated content can reveal different views of the same issues among people who write in different languages. The Meaning Extraction Method (MEM) was applied to compare posts from depression-related forums in Spanish and English [26]. MEM is used to discover the main topics in a corpus. A comparison of the resulting topics shows that English posts tend to use words that are more concrete and descriptive and the main topics are related to medicinal questions and concerns. Spanish posts use relatively more emotional words and the main topics are associated with sharing and disclosing information about relational concerns. Hecht and Gergle used the Explicit Semantic Analysis (ESA) algorithm to analyze pairs of terms from ten different Wikipedia language editions [14]. ESA indicates a score of semantic relatedness between two concepts. The findings reveal that "even when two language editions cover the same concept, they may describe that concept differently" [14]. For example, consider the pair "Germany" / "Saxony-Anhalt" (a state of Germany). In most languages this pair receives a high ESA score, but the algorithm detects no relation at all in Italian and Danish. This occurs because there are no articles that mention "Germany" and "Saxony-Anhalt" together in these two languages. Analyses of semantic networks and the salience of semantic concepts in articles about China in the

Chinese and English versions of Wikipedia found dissimilarities in the semantic content of these two versions [19]. Articles in the Chinese version are framed from the perspectives of respecting authority, emphasizing harmony and patriotism. Articles in English are written from the perspective of Western-societies' core value of democracy. The English version contains critical attitudes toward the authority of the Chinese government and the Communist Party in terms of human rights and territorial dispute. According to Jiang et al. cultures, values, interests, situations and emotions of different language groups can explain these dissimilarities. The latter studies provide evidence in support of what is known as the Sapir-Whorf hypothesis, which indicates that the structure of anyone's native language influences the world-views she or he will acquire as she or he learns the language [20]. Thus, speakers of different languages could think, perceive reality and organize the world around them in different ways [18]. Inspired by this hypothesis, we seek to study whether people who speak a different language hold different views of a data privacy scandal, such as the Cambridge Analytica case.

3   RESEARCH QUESTION

Given the relative lack of data privacy research in international settings [29], our project aims to investigate the potential of social media text written in different languages as a source to compare data privacy views worldwide. Beyond the Sapir-Whorf hypothesis (explained in Section 2), prior privacy research has argued that language and country of residence can relate to diverse perspectives on privacy. Smith et al. [29] noted that "many languages, including those in European countries (e.g Russian, French, Italian), do not have a word for privacy and have adopted the English word". Belanger and Crossler argued that "individuals from different countries can be expected to have different cultures, values and laws, which may result in differences in their perceptions of information privacy and its impacts" [3].

As the Cambridge Analytica scandal sparked worldwide conversations (in diverse languages) on Twitter about this particular misuse of user data, these online public communications are useful sources to contrast views on the scandal itself, its relation to data privacy, and its implications. To start addressing our ultimate research goal, this paper focuses on the following research question: "Which are the shared and unique topics that emerge from the Twitter activity in Spanish and English about the Cambridge Analytica data misuse scandal?"

4   DATA AND METHODS

To answer our research question, we used Tweepy[1], a Python library for accessing to the standard realtime streaming Twitter API. Using this library we were able to capture tweets that include hashtags or keywords related to the Cambridge Analytica scandal or data privacy, such as: "#CambridgeAnalytica", "#DeleteFacebook", "Zuckerberg" and "Facebook privacy". The standard realtime streaming Twitter API returns a random sample of all public tweets that match the search keywords. We collected tweets written in Spanish and English between April 1st and July 10th, 2018. Overall, we collected more than 7.4 million tweets written in English and more than 470, 000 tweets in Spanish (see Table 1). The English tweets

---

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

were generated by about 1.8 million unique Twitter accounts while the Spanish tweets were produced by approximately 220, 000 users. The difference between the number of tweets and users collected in English and Spanish suggests that English-speaking Twitter users tweeted more about this scandal using the selected keywords than Spanish-speaking users, although this may be explained by the greater volume of English tweets overall[2].

We cleaned our dataset in two ways. First, we removed all retweets to focus our study on original user opinions and avoid analyzing duplicates. This step downsized both datasets to approximately 20% of their original sizes. Second, we attempted to eliminate tweets generated by automated accounts so our study could indeed reflect people's opinions. Unfortunately, there is not yet an infallible mechanism to detect bots' activity on Twitter. We chose to use Botometer to identify potential bots [9]. Botometer implements a machine learning algorithm that has achieved high accuracy (0.94) in detecting both simple and sophisticated bots in prior work [34]. The algorithm has been trained to detect bots by analyzing Twitter accounts' metadata, their contacts' metadata, tweets' content and sentiment, network patterns, and activity time series. The result is a score that is based on how likely it is to be a bot. The score ranges from 0 to 1, where lower scores indicate that the account behaves like a human and higher scores signal bot-like behavior. Unfortunately, there is not yet agreement on a threshold that can reliably distinguish bots from humans. To define thresholds for our two datasets, we used the Ckmeans [35] algorithm to cluster the Botometer scores in each dataset into five groups, with the first cluster including the accounts with the lowest Botometer scores (more human-like) and the fifth group the users with the highest scores (more bot-like). We reasoned that the fourth and fifth clusters in each dataset were least likely to contain humans; therefore, we excluded them from our analysis. Given that the Botometer analysis is time-consuming, we analyzed only a sample of users. In this study, we focused on the users who contributed the highest number of tweets in our datasets. In the future we plan to analyze users who contribute less. Thus, we were able to classify 19, 478 accounts in the Spanish dataset (40.6%) and 74,021 (12.9%) accounts in the English dataset. Accounts with a Botometer score higher than 0.4745 were labeled as bots in the Spanish dataset. Those with a score higher than 0.4849 were considered bots in the English dataset. These users and their tweets were removed from our analysis. As a result, our final Spanish dataset includes 15, 531 users who tweeted 50, 559 times about the Cambridge Analytica scandal. The English dataset comprises 60, 491 accounts that generated 446, 462 tweets about it. Table 1 details these figures.

Table 1: Size of the Spanish and English datasets before and after data cleaning

Dataset	Spanish		English	
	#Tweets	#Users	#Tweets	#Users
Total	472,363	222,352	7,476,988	1,846,542
Without retweets	106,656	47,951	1,572,371	574,452
Most active	70,393	19,478	741,694	74,021
Humans	50,559	15,531	446,462	60,491

[2]https://www.statista.com/statistics/267129/most-used-languages-on-twitter/

To identify key topics in the resulting Spanish and English datasets, we used the Meaning Extraction Method (MEM) [8]. MEM is a topic modeling technique that can infer "what words are being used together, essentially resulting in a dictionary of word-to-category mappings from a collection of texts"[5]. After applying principal component analysis over this dictionary it is possible to identify words that can be grouped into themes or topics. This method has been identified as well-suited for cross-cultural and inter-language research [5]. MEM has been used to find themes in different contexts, such as: mental health [26, 38]; personality [6, 12] and values [37].

We employed the Meaning Extraction Helper (MEH) software [4], tool that can automate the majority of the MEM process [5]. The software contains a default list of Spanish and English stopwords; all these words were removed. Also, the tool allowed us to apply text segmentation by whitespace, conduct lemmatization and run Twitter-aware tokenization. To assist in lemmatization tasks, a conversion list was used to fix common misspellings (e.g "hieght" to "height") and convert "textisms" (e.g, "bf" to "boyfriend"). No stemming algorithm was used. We computed the frequency of each unigram as the percentage of tweets that contain it. The 300 most frequent unigrams were kept. We obtained a csv file with values of 1 and 0 indicating the corresponding unigrams' presence or absence, respectively, for each tweet.

Principal component factor analysis (PCA) was run over the MEH results. PCA was performed with varimax rotation to ensure that all resulting components are independent from each other. We conducted PCA with 5, 8, 11, 30 and 100 components. In both the Spanish and English datasets, 11 components gave the best results, with fit based upon diagonal scores of 0.55 and 0.9, respectively. This metric is a goodness of fit statistic, where values closer to 1 indicate better fit. The selected components accounted for 9% of the total variance of the Spanish data and 13% of the total variance of the English data.

To obtain the most representative words of each resulting component, we selected the words with factor loadings above 0.1, as recommended in [5]. The words were sorted according to their contribution to the component. Additionally, a python script was used to identify the top-30 tweets most related to each component. Using the most representative words and tweets by component, two authors examined and conceptualized the theme represented by each component, assigned a representative name and determined its relevance to our research question.

Finally, we used GeoNames API[3] to geo-locate all human-like accounts (see Table 1). Table 2 reports the proportion of tweets and users by the top-10 countries in each of our datasets. Most accounts could not be geo-located. The remaining accounts reveal that Spain and the US account for the majority of tweets and users in the Spanish and English datasets, respectively.

5 RESULTS

The words that clustered together to form coherent themes in the English and Spanish corpora are available online[4]. Tables 3 and 4 report the seven words with the highest loadings by component

[3]http://www.geonames.org/
[4]https://github.com/gonzalezf/LA-WEB-Paper

Table 2: Top ten most frequent user location in the English and Spanish datasets

Spanish			English		
Country	% tweets	% users	Country	% tweets	% users
not found	43.5%	46.05%	not found	41.6%	44.9%
Spain	17.5%	16.43%	U.S	33.6%	31.1%
Mexico	10.3%	9.90%	U.K	6.2%	6.9%
Venezuela	5.6%	3.56%	India	3.3%	3.0%
Argentina	5.4%	5.41%	Canada	2.4%	2.3%
Colombia	2.8%	2.98%	Australia	1.1%	1.4%
U.S	2.3%	2.40%	France	1.0%	0.6%
Chile	2.0%	2.35%	Germany	0.9%	0.7%
Peru	1.3%	1.41%	U.A.E	0.6%	0.3%
Ecuador	1.2%	1.25%	Netherlands	0.5%	0.4%
Brazil	0.9%	0.34%	Ireland	0.4%	0.4%

in the English and Spanish datasets, respectively. The tables also show the proportion explained (PE) by each of them according to the factor analysis. This number is proportional to the number of tweets associated with each theme. Table 5 presents the key themes in Twitter activity about the Cambridge Analytica scandal in Spanish and English. The themes are ordered according to their relevance to our research question.

Three key themes emerge in both languages. Spanish and English speakers talk about:

- "Cambridge Analytica's impact on political issues"
- "Mark Zuckerberg's Senate hearing in the USA," and
- "General Data Protection Regulation."

However, differences appear in how these themes are articulated. In regard to the first topic, the scandal's connection to Russia was much less relevant for Spanish speakers than for English speakers. English tweets focus on how Russia might have used Cambridge Analytica to intervene in the 2016 US elections and the UK Brexit campaign. For example, this component includes the following tweet: *"@ianbremmer @billmaher @RealTimers You want to know how Brexit happened and Trump got to win? Cambridge Analytica, Bannon, Mercers and the Bot Farms in Russia. They started testing MAGA, Build the Wall, Lock Her Up, Anti-Muslim sentiment, Anti-Immigrant propaganda. Putin worked hard at it since 2012"*. On the other hand, the token *Russia* does not appear as a representative word of this theme in the Spanish dataset. Instead, Spanish tweets are centered on Cambridge Analytica's closure as a result of the scandal. This behavior could be explained by the users' country of residence and its closeness to a salient political issue related to Cambridge Analytica. The US accounts for the largest share of users who report their location in our English dataset (see Table 2). It is well known that many people from this country have apprehensions about Russia since the Cold War. Furthermore, recent research has found evidence that Americans express "continued mistrust of Russia and a majority think Russia tried to interfere in the 2016 election" [27]. We believe that this is a plausible reason to explain this difference between the English and Spanish tweets.

While both datasets include a topic about "Mark Zuckerberg's Senate hearing in the US", their tweets' verb tenses differ. English-speaking users tend to tweet about this topic in future or present tense. These tweets reflect either certain level of anticipation of the event or live reports on how the event was unfolding. An example of these tweets is: *"Facebook CEO Mark Zuckerberg will testify in front of a joint hearing of the Senate Judiciary and Commerce Committees today. Senators will demand answers from Zuckerberg about Facebooks failure to protect up to 87 million users' private information https://t.co/9qpdqxjM7F https://t.co/aLlZtMMVCC"*. On the other hand, Spanish-speaking users are more likely to discuss this issue in past tense. They comment on sentences that Zuckerberg said during the hearing, putting special emphasis on the moment when he assumed responsibility for what has happened. For instance, translated Spanish tweets state: *"From #Whoknows to #ThroughMyFault, the change of attitude of #MarkZuckerberg. 'It was my mistake, and I'm sorry, I started #Facebook and I'm responsible for what happens here': Mark Zuckerberg before the Congress of #EEUU https://t.co/UX2QwT7pFw"*, and *"Zuckerberg takes full blame for the abuse of Cambridge Analytica before the US Senate: 'It was my mistake, and I'm sorry' https://t.co/VjH9ocdLcB"*. The difference in tenses (English future/present and Spanish past) may be explained by a delay in news reporting (volume of the Spanish tweets in our dataset relating to this particular topic tended to peak about 24 hours after the peak occurred in the English tweets) and translation to another language as the event occurred in an English-speaking country.

In the case of the "General Data Protection Regulation" (GDPR) theme, a tendency to attribute responsibility to companies for not attempting to comply with the GDPR was present in the English data, but not in the Spanish tweets. The English dataset includes accusatory tweets to companies such as Facebook and Google for trying to dodge GDPR rules, e.g. *"Facebook and Google are pushing users to share private information by offering invasive and limited default options despite new EU data protection laws aimed at giving users more control and choice https://t.co/x0srYwZeUz"*, *"If a new European personal data regulation (aka #GDPR) went into effect tomorrow, almost 1.9bn #Facebook users around the world would be protected by it. The online social network is making changes that ensure the number will be much smaller.https://t.co/UXonA0mTCs #privacy https://t.co/gqFaMYq9Su"*, and *"Facebook moves billions of international user accounts to California to avoid European privacy law https://t.co/nC2ccf2ClR"*. On the other hand, Spanish-speaking users tweet about GDPR to describe it and inform local companies how to prepare for it. Some translated tweets belonging to this component are: *" General #Data Protection Regulation (#GPDR) is a new #law of the European Union and it will enter into force on May 25th. Do you know what it is? Do you know its characteristics? Is your #company ready? Get up to date by clicking on the following link! https://t.co/AeKPR4w6Q9 https://t.co/rdlzbrAU1n"*, and *"On May 25th, the General Protection #Data Regulation (GDPR) of the EU, one of the most modern regulations regarding personal data use by companies and institutions, begins to be applied. There will be some repercussions in #Chile. https://t.co/XNbmALUcdD"*

Three other emerging themes are unique to Spanish speakers and are relevant to our research question as they all refer to the founder of Facebook and its role in different aspects of the scandal. These topics are:

- "Zuckerberg in front of the European Parliament"
- "Opinions about Mark Zuckerberg," and

Table 3: Terms by themes in the Spanish dataset

#ID	PE (%)	Words with the highest loadings by component						
		word1	word2	word3	word4	word5	word6	word7
S1	13%	animals	lottery	predict	live	program	listen	win
S2	11%	cambridge	analytica	million	user	scandal	data	affect
S3	10%	congress	error	zuckerberg	sorry	mark	usa	senate
S4	9%	rgpd	protection	gdpr	data	regulation	privacy	dataprotection
S5	9%	marketing	digitalmarketing	analytic	publicity	digital	google	youtube
S6	9%	press	like	followers	platform	follow-us	come	digital
S7	9%	red	social	twitter	instagram	youtube	facebook	follow-us
S8	8%	parliament	european	zuckerberg	mark	ask	hearing	sorry
S9	8%	stop	ask	join	people	red	create	senate
S10	8%	communitymanager	socialmedia	blog	socialnetworks	news	work	facebook
S11	7%	message	messenger	user	send	million	remove	delete

Table 4: Terms by themes in the English dataset

#ID	PE (%)	Words with the highest loadings by component						
		word1	word2	word3	word4	word5	word6	word7
E1	24%	newyorkcity	newyork	nyc	ny	career	code	hire
E2	12%	rsi	btc	signal	min	eth	bitcoin	crypto
E3	10%	machinelearn	deeplearn	artificialintelligence	ml	robotic	dl	ai
E4	10%	chatbot	infosec	databreach	hack	cybersecurity	crypto	blockchain
E5	9%	iot	iiot	smartcity	digitaltransformation	innovation	infographic	startup
E6	7%	cambridge	analytica	election	campaign	voter	vote	brexit
E7	7%	data	user	facebook	privacy	access	personal	law
E8	6%	zuckerberg	mark	testify	ceo	congress	committee	senate
E9	6%	trump	president	donald	white	house	democrat	america
E10	5%	social	media	twitter	instagram	censorship	conservative	facebook
E11	5%	late	daily	thanks	bigdata	remove	social	ai

Table 5: Themes in the Spanish and English datasets

Spanish		English	
Theme	#ID	Theme	#ID
Cambridge Analytica's impact on political issues	S2	Cambridge Analytica's impact on political issues	E6
Mark Zuckerberg's Senate hearing in the US	S3	Mark Zuckerberg's Senate hearing in the US	E8
General Data Protection Regulation	S4	General Data Protection Regulation	E7
Zuckerberg in front of the European Parliament	S8	Facts and opinions about Donald Trump	E9
Opinions about Mark Zuckerberg	S9	Stop censorship on social media	E10
Facebook deletes Zuckerberg's private messages	S11	Cryptocurrencies	E2
Digital marketing	S5	Artificial intelligence	E3
Promoting subscription to social platforms	S7	Blockchain	E4
Promoting likes in social platforms	S6	Internet of things	E5
Lottery results	S1	News about privacy on social media	E11
News and random facts	S10	Hiring tech jobs in New York	E1

- "Facebook deletes Mark Zuckerberg's private messages."

The first topic focused on Zuckerberg's laments for the situation, e.g. *"Mark Zuckerberg apologized to the European Parliament for the data breach. The founder of Facebook acknowledged on Tuesday that the tools of the social network were used 'to do harm'. https://t.co/aeDUVViUhq"* We note that this theme only emerges in the Spanish tweets. Again, we attribute this distinction to the users' country of residence. Both datasets include users whose self-reported location is in Europe; however, they are the majority only in the Spanish dataset where Spain is associated with more tweets than any other country (see Table 2). Thus, the hearing in the EU Parliament is much more prominent in the Spanish dataset.

The second topic contains supporting and accusatory tweets for Mark Zuckerberg in relation to the Cambridge Analytica scandal. This theme includes tweets such as the following: *"what they do not forgive to Zuckerberg is that Facebook has unwittingly helped Trump triumph (which I do not think), because when Facebook censored pages of right-wing groups for nothing (the same Mark Zuckerberg has said that he worries about the leftist prejudice of his staff), there they said nothing."*

The third theme groups together tweets about the option to delete private direct messages on Facebook. This functionality apparently was available to Mark Zuckerberg in the past while no other user could use it. Spanish speakers tweet about the special privileges of Zuckerberg as a Facebook user and about the possibility that such functionality could become available to everyone. Example translated tweets are: *"Although you can not delete your messages sent from another person's inbox in Messenger, Facebook has done so in the case of Mark Zuckerberg and other company executives. https://t.co/PkJByaesdT"*, and *"It is clear that Mark Zuckerberg is the only one who can have the God Mode of Facebook because only he can erase messages sent through Facebook Messenger, a feature that will be released for everyone after this scandal."*

Two themes are unique to English and are considered slightly relevant to our research question. They are conceptualized as:

- "Stop censorship on social media," and
- "News about privacy on social media."

First, English-speaking users protest about censorship on social media companies such as Facebook, Twitter, Instagram and Google. They demand freedom of speech. This component includes the following tweet: *"@facebook Social media censorship is tearing out our tongues! Is Facebook a neutral public forum? STOP CENSORSHIP! Demand an #InternetBillOfRights #IBOR #MAGA #InvasionOfPrivacy #HumanRights #DeleteFacebook #Censorship #Twitter #Censorship #Google #FaceBook #Instagram @realDonaldTrump"*. The next topic includes informative tweets (usually from electronic newspapers) that relate to social media and privacy, e.g. *"The latest The GDPR and Data Protection Daily! https://t.co/GZBscXkxPl Thanks to @simpledatainc @Colt_Technology @content_app #bigdata"*.

Other key themes arise from both datasets, but we consider them not sufficiently relevant to our research question. In the Spanish dataset, these themes are: "Digital marketing", "Promoting likes in social platforms", "Promoting subscription to social platforms", "Lottery results", and "News and random facts". In English, there are the following topics: "Cryptocurrencies", "Artificial Intelligence", "Blockchain", "Internet of Things", "Facts and opinions about Donald Trump," and "Hiring tech jobs in New York."

6 DISCUSSION

Our study of tweets in Spanish and English about the Cambridge Analytica data misuse scandal allowed us to conduct an inter-language comparison of their main emerging themes. Out of eleven topics, three are common to both languages. However, they present meaningful differences in their articulation. Three other relevant themes are unique to the Spanish data and two relevant topics only arise from the English tweets. These findings show the potential of our algorithm-based approach for inter-language comparisons of tweets

to identify similarities and (nuanced) differences on privacy-related views across people from different countries.

We discuss here two underlying patterns that could explain the distinctions we found between Spanish and English speakers.

First, we observe a tendency for a local perspective to appear in tweets about this data privacy scandal. Tweets in English provide elaborate rationales behind the Cambridge Analytica's impact on political issues. They often relate them to Russia's actions. This is a common argument among Americans [27], who are also the most active contributors of tweets in our English dataset. This kind of rationale is much less visible in tweets in the Spanish dataset, where most users are not located in the US. Furthermore, there is a large component of tweets about Mark Zuckerberg's hearing in front of the US Senate in both languages; however, only the Spanish corpus has a theme related to Zuckerberg's audience in the European parliament. Again, we relate this finding to the users' country of residence. Our Spanish dataset consists primarily of European users. This could explain why this event became more salient only in this language. Another finding that can be explained by a local perspective is the demand for freedom of speech in the English tweets. This theme does not emerge at all in the Spanish data. Free speech is a central value in the US [2], but it is not as prominent in Spanish-speaking countries.

Second, English tweets often attribute responsibilities to governments and organizations while Spanish tweets tend to question individual actions. For example, English-speaking users tend to tweet accusatory statements about big companies such as Google and Facebook dodging GDPR. On the other hand, these large companies are very rarely questioned about their acts in the Spanish data. Instead, it is possible to find many tweets that support Mark Zuckerberg and blame Facebook users for being irresponsible by failing to read Facebook's terms of service and neglecting to protect their personal data.

Research on cross-cultural differences could provide a framework to further explore the second pattern. This research argues that national cultures could be characterized by their scores on a small number of dimensions [24]. Hofstede's cultural dimensions [15] have been widely used to study the relationship between people's culture and technology [24]. We hypothesize that one of these dimensions, "power distance", could explain differences between Spanish and English speakers on responsibility attribution. This dimension is defined as: "the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally" [16]. In countries with a high level of power distance, people tend to accept hierarchies without further justification and authority is hardly questioned. Most Anglo and Nordic countries are characterized by small power distance scores. The opposite was found in Latin American countries. These nations often score high in power distance [15]. This pattern could explain why many tweets that question authorities (e.g., governments and companies) occur in English while they are much less present in Spanish. While this is a plausible path of reasoning, cross-cultural research has its own limitations [31, 32]; therefore, additional work is needed to confirm or refute this explanation.

In both the English and the Spanish dataset we found tweets from spammers and self-promoters' accounts. These accounts tend

to use hashtags relevant to popular online discussions (e.g: #Facebook, #privacy) to promote their conversations and possibly share malicious links. These tweets account for the themes that were considered irrelevant in our analysis.

6.1 Limitations and future work

As in any study, our research has limitations that need to be taken into consideration. We collected data through the standard streaming Twitter API and by using specific hashtags and keywords. Thus, we only had access to a small sample of all the tweets about the scandal. Nevertheless, studies have shown that this API is highly correlated with the full Twitter stream [23]. While we attempted to identify and remove bot activity from our data, and the threshold we defined lies between the recommended range [0.43-0.49] [34], our dataset contains both false positives and negatives, which could affect the results. Also, difficulties interpreting MEM results could occur due to the lack of contextual information such as "information about valence and timing that is obscured in the analysis" [11]. Further qualitative analysis may help with future interpretation.

Future work includes investigating the English and Spanish corpora in order to strengthen the analysis of shared themes. This could be done, for example, by incorporating measurements of similarity between texts. Using bigrams and trigrams as inputs of the MEM analysis could help to contextualize the topic components. Also, it could be interesting to compare MEM results with other topic modeling techniques such as LDA or LDA2Vec. Furthermore, exploring the data over the time dimension should be considered because it may allow the detection of topic changes over time.

7 CONCLUSIONS

We proposed a methodology to conduct an inter-language study of Twitter activity about a major data privacy leakage: the Cambridge Analytica scandal. We collected tweets about it in English and Spanish during 100 days, processed them to remove tweets generated by bots, and conducted MEM to identify the key themes in each language. We conceptualized the themes and compared them. We identified five topics that were unique to only one of the languages. We also found three common topics: "Cambridge Analytica's impact on political issues," "Mark Zuckerberg's Senate hearing in the USA," and "General Data Protection Regulation." Nevertheless, we detected dissimilarities in how these topics were formulated in each language. We proposed two plausible patterns that may explain these differences. First, reactions to a data privacy scandal reflect the users' local perspectives. Second, there is a tendency for English speakers to assign responsibilities to governments and organizations while Spanish speakers tend to attribute them to people.

The contributions of our work and findings are twofold. First, we proposed a method to leverage social media data to investigate reactions to a data privacy scandal in two languages that cover a multi-national audience. Second, we found inter-language differences on privacy-related views that go beyond measuring levels of privacy concerns, as defined in widely-adopted surveys such as [30]. We expect that future work can deepen the understanding of these differences through diverse methods of research.

REFERENCES

[1] Adam Acar. 2013. Culture and social media usage: Analysis of Japanese Twitter users. *International Journal of Electronic Commerce Studies* 4, 1 (2013), 21–32.
[2] Mauricio J Alvarez and Markus Kemmelmeier. 2018. Free Speech as a Cultural Value in the United States. *Journal of Social and Political Psychology* 5, 2 (2018), 707–735.
[3] France Bélanger and Robert E Crossler. 2011. Privacy in the digital age: a review of information privacy research in information systems. *MIS quarterly* 35, 4 (2011), 1017–1042.
[4] RL Boyd. 2014. MEH: Meaning Extraction Helper (Version 1.0. 6)[Software].
[5] Ryan L Boyd. 2017. Psychological text analysis in the digital humanities. In *Data analytics in digital humanities*. Springer, 161–189.
[6] Ryan L Boyd and James W Pennebaker. 2015. Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological science* 26, 5 (2015), 570–582.
[7] Kyungsub Stephen Choi, Il Im, and Gert Jan Hofstede. 2016. A cross-cultural comparative analysis of small group collaboration using mobile twitter. *Computers in Human Behavior* 65 (2016), 308–318.
[8] Cindy K Chung and James W Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality* 42, 1 (2008), 96–132.
[9] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 273–274.
[10] Tamara Dinev, Massimo Bellotto, Paul Hart, Vincenzo Russo, Ilaria Serra, and Christian Colautti. 2006. Privacy calculus model in e-commerce–a study of Italy and the United States. *European Journal of Information Systems* 15, 4 (2006), 389–402.
[11] Marilyn R Fitzpatrick and Calli Renee Armstrong. 2010. Beyond the tip of the iceberg: Exploring the potential of the meaning extraction method and aftercare e-mail themes. *Psychotherapy research* 20, 1 (2010), 86–89.
[12] Eden M Griffin and Karen L Fingerman. 2018. Online Dating Profile Content of Older Adults Seeking Same-and Cross-Sex Relationships. *Journal of GLBT Family Studies* 14, 5 (2018), 446–466.
[13] Michael M Harris, Greet Van Hoye, and Filip Lievens. 2003. Privacy and attitudes towards internet-based selection systems: a cross-cultural comparison. *International Journal of Selection and Assessment* 11, 2-3 (2003), 230–236.
[14] Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 291–300.
[15] Geert Hofstede. 1983. National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization* 13, 1-2 (1983), 46–74.
[16] Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.
[17] Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language Matters In Twitter: A Large Scale Study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 518–521.
[18] Basel Al-Sheikh Hussein. 2012. The Sapir-Whorf Hypothesis Today. *Theory and Practice in Language Studies* 2, 3 (2012), 642–646.
[19] Ke Jiang, Grace A. Benefield, Junfei Yang, and George Barnett. 2017. Mapping Articles on China in Wikipedia An Inter-Language Semantic Network Analysis. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. IEEE.
[20] Paul Kay and Willett Kempton. 1984. What is the Sapir-Whorf hypothesis? *American anthropologist* 86, 1 (1984), 65–79.
[21] Naz Kaya and Margaret J Weber. 2003. Cross-cultural differences in the perception of crowding and privacy regulation: American and Turkish students. *Journal of environmental psychology* 23, 3 (2003), 301–309.
[22] Jean-Rémi Lapaire. 2018. Why content matters. Zuckerberg, Vox Media and the Cambridge Analytica data leak. *ANTARES: Letras e Humanidades* 10, 20 (2018), 88–110.
[23] Kalev Leetaru. 2019. Is Twitter's Spritzer Stream Really A Nearly Perfect 1% Sample Of Its Firehose?
[24] Dorothy E Leidner and Timothy Kayworth. 2006. A review of culture in information systems research: Toward a theory of information technology culture conflict. *MIS quarterly* 30, 2 (2006), 357–399.
[25] Maya Mirchandani. 2018. To delete or not to# Delete Facebook, that is the question.
[26] Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*.
[27] Dina Smeltz and Lily Wojtowicz. 2018. Public Opinion on US-Russian Relations in the Aftermath of the 2016 Election. *SAIS Review of International Affairs* 38, 1 (2018), 17–26.

[28] H Jeff Smith. 2001. Information privacy and marketing: What the US should (and shouldn't) learn from Europe. *California Management Review* 43, 2 (2001), 8–33.

[29] H Jeff Smith, Tamara Dinev, and Heng Xu. 2011. Information privacy research: an interdisciplinary review. *MIS quarterly* 35, 4 (2011), 989–1016.

[30] H Jeff Smith, Sandra J Milberg, and Sandra J Burke. 1996. Information privacy: measuring individuals' concerns about organizational practices. *MIS quarterly* (1996), 167–196.

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

[32] Rosalie L Tung and Alain Verbeke. 2010. Beyond Hofstede and GLOBE: Improving the quality of cross-cultural research. *Journal of International Business Studies* 41, 8 (01 Oct 2010), 1259–1274.

[33] Blase Ur and Yang Wang. 2013. A cross-cultural framework for protecting user privacy in online social media. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 755–762.

[34] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).

[35] Haizhou Wang and Mingzhou Song. 2011. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R journal* 3, 2 (2011), 29.

[36] Yuanyuan Wang, Muhammad Syafiq Mohd Pozi, Yukiko Kawai, Adam Jatowt, and Toyokazu Akiyama. 2017. Exploring Cross-cultural Crowd Sentiments on Twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 321–322.

[37] Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 143–152.

[38] Markus Wolf, Cindy K Chung, and Hans Kordy. 2010. Inpatient treatment to online aftercare: e-mailing themes as a function of therapeutic outcomes. *Psychotherapy Research* 20, 1 (2010), 71–85.