Cross-Cultural Studies Using Social Networks Data

Issa Annamoradnejad, MohammadAmin Fazli, Jafar Habibi, and Sadjad Tavakoli

***Abstract*— With the widespread access of people to the Internet and the increasing usage of social networks in all nations, social networks have become a new source to study cultural similarities and differences. We identified major issues in traditional methods of data collection in cross-cultural studies: difficulty in access to people from many nations, limited number of samples, negative effects of translation, positive self-enhancement illusion, and a few unreported problems. These issues are either causing difficulty to perform a cross-cultural study or have negative impacts on the validity of the final results. In this paper, we propose a framework that aims to calculate cultural distance among several countries using the information and cultural features extracted from social networks. To this aim, the framework estimates the distribution of news-oriented tweets for each nation and computes the cultural distance from these sets of distributions. Based on a sample composed of more than 17 million tweets from late 2017, our framework calculated cultural distance between 22 countries. Our results show a positive correlation between cultural distances computed by our framework and distances computed by Hofstede's cultural scores and also identified connections between some of the cultural features.**

***Index Terms*— Cross-cultural study, cultural distance, social networks, social network analysis, Twitter**

I. Introduction

Cross-cultural studies are the science of using data from different societies to test hypotheses about human behavior and culture. In order to perform a cross-cultural study, researchers need to aggregate and compare data from multiple countries which is suggested to be the main extra problem for this type of research compared to others [1]. We thoroughly investigated several works that studied culture in multiple nations and identified a few major issues in the phase of data collection. These issues are either causing difficulties to perform a cross-cultural study or have negative impacts on the validity of the final results. Major points on these issues are provided in the next paragraphs.

In cross-cultural studies, it is always a requirement to collect data from multiple nations. In most cases, this is accomplished by spreading questionnaires and aggregating the individual results, which requires access of researchers to several people from the selected nations. For example, in the study of value hierarchies of individuals in different cultures [2], Schwartz and Bardi collected the data from 63 nations using a survey that was translated into 39 languages. In another study, Terracciano *et al.* [3] studied relationships between national character and mean personality traits levels by collecting the data from 49 cultures/subcultures. Hofstede [4] questioned the employees of IBM, as a multinational company, to collect initial data for his prominent work on cultural dimensions. Table I contains more information on major cross-cultural studies that collected original data from multiple countries.

The limited number of participants is another major issue with the traditional methods of data collection in cross-cultural studies. As given in Table I, past studies used less than 1200 people per country in their research. This low number of participants raises the question of validity on final results. In fact, in some cases, researchers excluded countries from the rest of their research for this reason. Furthermore, participants sometimes prefer not to completely fill the questionnaire, leading to the removal of those answers for failing to comply with the minimum requirement of the questionnaire. As an example of these problems, in the study of national culture and values of organizational employees [5], from 10 993 of business organizational employees who answered the questionnaire, 9920 items satisfied the necessary completeness. In addition, because of low number of respondents, the results of 12 nations were excluded from the study.

In many cross-cultural research studies (such as [2], [3]), researchers translated questionnaires in order to reach more people. The negative impacts of translation into the final output have been studied by multiple studies [6], [7], where it is suggested that the careful act of translating questionnaires into multiple languages takes extra time and energy from researchers. In the study of sex difference in big five personality traits [8], a nation was completely excluded from the research because of errors in translation. As another example, in the study of national culture and values of organizational employees [5], some of the collected data were excluded from the article for the reason of inaccuracies in translation. As for time consumption of this process, Schwartz and Bardi [2] translated their questionnaire into 39 different languages, McCrae and Terracciano [9] translated into 27 languages, and in the study of patterns of geographic distribution of big five personality traits, Schmitt *et al.* [10] translated their questionnaire into 29 different languages (except for five countries that used survey in English) alongside descriptions for some phrases or terms that could have been misleading or confusing for readers.

In addition to these major issues, there are some unreported problems to the traditional methods of data collection. Psychologists mention "positive self-enhancement illusions" as a general behavior of a normal human, meaning that most people have a positive illusion about themselves [11], [12]. In our case, this suggests that because of a tendency of people to self-enhancement, participants evaluate themselves more positively

Manuscript received December 16, 2018; revised May 9, 2019; accepted May 23, 2019. *(Corresponding author: MohammadAmin Fazli.)*

The authors are with the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran (e-mail: fazli@sharif.edu).

Digital Object Identifier 10.1109/TCSS.2019.2919666

TABLE I
SUMMARY OF DATA COLLECTION IN RELATED CROSS-CULTURAL WORKS

Subject	# Countries	Data summary
The rotter locus of control scale in 43 countries: a test of cultural relativity[15]	43	Results based on 9,140 participants from 43 countries. They were selected out of 10,993 participants from 58 countries.
Culture in the cockpit do Hofstede's dimensions replicate?[16]	19	Results based on 9,417 pilots from 26 airlines in 19 countries.
Cultural influences on the relation between pleasant emotions and unpleasant emotions[17]	38	Results based on 5,886 participants (3,653 women, 2,233 men) from 38 countries. They were selected out of 6,780 college student from 40 countries, excluding others because of missing or extreme values.
Meanings of Basic Values for Women and Men: A Cross-Cultural Analysis[18]	-	Results based on 11,244 participants from 8 cultural regions.
Culture-level dimensions of social axioms and their correlates across 41 cultures[19]	41	Results based on 7,654 university students from 41 cultures and 2,252 adults from 13 cultures.
Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies[20]	59	Results based on 17,370 middle manager participants from 951 companies in 59 countries / 62 cultures.
Personality Profiles of Cultures: Aggregate Personality Traits[9]	51	Results based on 12,156 college students from 51 cultures.
Universal Intracultural and Intercultural Dimensions of the Recalled Frequency of Emotional Experience[21]	48	Results based on 9,291 participants (5,611 women and 3,680 men from 48 countries.
The geographic distribution of big five personality traits[10]	58	Results based on 17,408 participants from 58 countries (Equal number of men and women from each country). They were selected out of 17,837 participants, excluding others because of missing or incomplete values.
How selfish are self-expression values? A civicness test[22]	48	Results based on 56,000 participants from 48 countries.
Why Can't a Man Be More Like a Woman? Sex Difference in Big Five Personality Traits Across 55 Cultures[8]	55	Results based on 17,637 participants from 55 countries. One nation was eliminated because of translation errors.

than others and there is always a chance of biased results in the self-reported data. Mistakes of respondents in filling forms and filling them without reading or in a random order can influence the results of studies without being noted by researchers. In addition, language tone, words, and sentences that are used in questions or during the interview can potentially impact the results.

In this study, we aim to analyze different aspects of using social networks as a new source of data collection in cross-cultural studies. To this aim, we propose a framework to examine cultural differences and similarities across multiple nations using the data gathered from social networks. We will use this framework to compute cultural distance among 22 countries by evaluating the distribution of tweets over six news categories. We developed an instance of our framework as an automated system that tracks the amount of attention that each one of these news categories has received from the people of a country. Then, it uses the distribution of news-oriented tweets as a cultural feature and calculates cultural distances from these sets of distributions. In order to classify tweets into news categories, we used a supervised text classification technique, which with respect to the classification challenges of this study performed with an acceptable precision.

In regard to the above-mentioned issues, we found our framework to be extremely helpful. By collecting data from social networks, there was no requirement to find participants from different nations using time- and energy-consuming methods. The number of collected items in our research in comparison to previous cross-cultural studies exceeded by a factor of 100 or more. Since we extracted cultural information by analyzing users' actions in social networks and not from direct questions, issues related to form translations and self-enhancement illusion are completely out of scope. While we translated all the input data to a target language, this translation was merely on the content in order to extract features and not on the questions, therefore, had no impact on users' behaviors or answers. Furthermore, since users of social networks are acting in their normal state, we have no reason to believe that positive self-enhancement illusion has an impact on our method.

These being said, we had a few new difficulties or challenges in our method of data collection. Our main difficulty in accessing people from different nations was the problem of detecting country of a Twitter user. Although it is unlikely to correctly detect the origin country of every user, we handled this problem by training a previously suggested classification model. In addition, to extract a specific cultural feature from social networks data, it is required for researchers to find or design a computer model. There are many models and methods in the field of big data, machine learning, and social networks analysis and mining, which are dedicated to extracting features from users' graph and actions.

In the remainder of this paper, we discuss a brief review of relevant literature in Section II. The technical framework is introduced in Section III. We examine our technical framework in an experimental setup in Section IV and evaluate our results in Section V. Concluding remarks are provided in Section VI.

II. RELATED WORKS

Many studies discuss human behavior across multiple cultures. While some of these studies preferred to use previously published data, the rest collected new data sets through traditional methods of data collection. As we discussed this in Section I, these traditional methods of data collection are prone to human errors and usually have characteristics that make them difficult to perform on a large scale. In this section, we are going to review data collection in several acclaimed cross-cultural studies and explore two works based

on online resources. At the last part of this section, we will review the literature on a few technical components of our framework.

A. Data in Cross-Cultural Studies

We investigated cross-cultural studies that generated new data sets. Table I contains the summary of our analysis on these works.

Cross-cultural studies have been highly influenced by the works of Hofstede [4], who proposed that a culture can be defined by six dimensions. Individualism versus collectivism. Uncertainty avoidance. Power distance (strength of social hierarchy). Masculinity versus femininity (task orientation versus person orientation). Long-term orientation. Indulgence versus self-restraint.

Hofstede investigated the aspects of these dimensions for 3 decades and calculated values of cultural scores for more than 90 countries by interviewing employees of IBM as a multinational company. While there have been a few debates and objections to his methods and choice of dimensions, many of the further cross-cultural studies compared their findings with Hofstede's cultural scores [13], [14].

B. Online Sources of Data Collection

While there have been several works on culture and society using social networks data (such as [23]–[25]), we only found two cross-cultural studies that collected data from online sources. There exist other cultural studies by mining social networks, but they were dedicated to the study of one culture and not to examine multiple cultures.

Callahan [13] inspected cultural differences and similarities between eight countries in the design of university websites, specifically their homepages. The study manually investigated several hypotheses about university websites. These hypotheses were based on two criteria: organization (general contents of the web pages) and graphical design. To evaluate results, the study correlated the frequency of interface elements of the web pages with Hofstede's cultural dimensions.

Park *et al.* [14] examined the usage of emoticons on Twitter from the users of 78 countries. They also used correlation analysis to evaluate their findings with Hofstede's cultural dimensions. The study found a positive correlation between the individualism-collectivism cultural dimension of Hofstede model and people's use of mouth-oriented (vs. eye-oriented) emoticons.

C. Related Works on Technical Issues

In our technical framework, we will use a supervised text classification algorithm to determine news orientation of a given tweet. The main challenges of the text classification process in our problem are as follows.

1) Tweets are very short (maximum 140/280 characters), therefore finding a meaningful relationship between the words of a tweet is almost impossible. This lack of relationship is enough for many of the classification algorithms to be unusable for this study.
2) Tweets are usually informal, both in structure and vocabulary. In addition, it is a common practice for Twitter users to use abbreviations to fit their content into the character limit (first challenge).
3) The collected tweets in this framework are multilingual; therefore, the precision of the classification algorithm can be expected to be lower than the cases of a single language.
4) Tweets are collected randomly without any kind of content filtering, and therefore, there is no context in order to simplify the text classification algorithm.

We had difficulty in finding a suitable text classification technique that would satisfy the text classification challenges of our framework. While there have been many approaches to classify texts in general, only a small fraction of them was dedicated to the classification of tweets and short texts (such as [26] and [27]). Zhang *et al.* [28] showed that the methods for classifying ordinary verbal communication, such as email and forum discussion, do not fit for short texts.

To the best of our knowledge, most of the recent works in short text classification are based on techniques designed for a specific language; for example, by using speech acts, stemming, clue words [29]–[31], n-gram and k-nearest neighbors [32], [33]. Thus, we were unable to find any method to handle multiple languages with one data set.

III. Framework

In this section, we are going to explain the technical details of our framework. In its high-level abstraction form, this framework collects data from social networks, classifies them by country, extracts meaningful features from data for each country, and finally, calculates cultural distance using those features. In the remainder of this section, we explain an instance of this framework.

First, by exploring major news broadcasting websites (Section III-A), we selected six broad news categories to be used as the set of classes in the text classification algorithm (Section III-E). Then, we trained the classification model by using news articles and tweets from news agency websites (Section III-B). By collecting a data set of tweets, we created a classification model to identify the country of a given Twitter user (Section III-C). We chose a target language to translate all items of the training and test data set (Section III-D). As a result, the training and test data set only consists of the translated version of the collected news articles to the target language.

By applying the classification model to all inputs (Section III-E), we aggregate the results of the classification algorithm for each country in order to obtain the distribution of tweets over the selected categories. Finally, we use these sets of distributions to compute the cultural distance between the selected nations (Section III-F).

It is worthy to note that any section of the framework, such as the classification algorithm, the set of news categories, or the cultural distance algorithm, can be replaced with other alternatives.

Further details of the method are discussed in Sections III-A–III-F.

TABLE II
MAIN NEWS CATEGORIES SELECTED FOR THIS STUDY

News Category	The Guardian	CNN	Reuters
Politics	Politics	Politics	Politics, World
Economics	Business	Money	Business, Markets
Art & Entertainment	Culture	Entertainment	Entertainment
Sports	Sport, Football	Sport	Sports
Science & Health	Science	Health	Health
Technology	Tech	Tech	Tech

TABLE III
NUMBER OF TWEETS FOR EACH COUNTRY AFTER APPLYING OUR DICTIONARY

Country	Original tweets	All tweets (retweets included)	Percentage of original tweets
Arab Countries*	592843	1074961	0.552
Argentina	7981	17159	0.465
Australia	20468	37750	0.542
Brazil	123751	237992	0.52
Canada	36640	81992	0.447
China	4286	11634	0.368
Colombia	28765	56639	0.508
France	76073	183593	0.414
Germany	28232	46476	0.607
India	73119	164738	0.444
Indonesia	78268	154296	0.507
Iran	12226	19461	0.628
Italy	13209	26822	0.492
Japan	2353109	3309549	0.711
Korea	289015	772588	0.374
Philippines	55861	105964	0.527
Russia	36722	59418	0.618
Spain	35434	74950	0.473
Thailand	68458	544220	0.126
Turkey	81240	195322	0.416
UK	158606	301037	0.527
USA	779829	1720510	0.453
Average	225188	418048.7	0.539
Sum	4954135	9197071	0.539

*: We used one row for all Arab countries in this article; since (1) Hofstede provided one aggregated result for all of them and (2) because of the challenges we had in separating their tweets.

A. News Categories

By analyzing top-level news categories used by some of the biggest worldwide news broadcasting websites, we selected a shortlist of most common and broad news categories. The list includes six subjects and we will use them as the set of classes in the text classification algorithm. Each one of these items is displayed as a separate news genre on the top level menu of several news broadcasting websites.

The selected categories for the classification algorithm and their associated labels in three well-known news websites are listed in Table II. It should be noted that in some cases, different websites used different or multiple labels to refer to the same category in our list. For example, since *business*, *markets*, and *money* sections are close in their nature, we created an *Economics* category to include topics related to all of these labels.

B. Training Data

In supervised text classification methods, the training data have a key role in the accuracy of the algorithm, and its quality and quantity have great consequences on the classifier's output. Based on our observation on studies that used a classification algorithm, the majority of researchers manually classified a fraction of the input data and used them as the training data set. In some cases, methods or resources have been proposed to collect and generate the training data set, automatically and without the need for manual intervention. As an example, Hu *et al.* [34] proposed the online encyclopedias as a good source to automatically collect and generate training data set.

In this study, we propose articles and tweets from news broadcasting websites as suitable sources for the training data set of this framework. They are ideal solutions to generate the training data set for several reasons, including: their open-access policy, up-to-date data, and previously classified texts (for example, the collected articles from these websites have tags and labels, which makes it easy to automatically assign a news category).

It should be noted that by creating an automated tool to train the classification model (by parsing the web pages of the recent news articles, really simple syndication feed of websites as an easier solution or their twitter accounts), our framework can be developed as a fully automated process to calculate the cultural distance between several nations.

TABLE IV
DISTRIBUTION OF TWEETS OVER NEWS CATEGORIES BY COUNTRY

Country	Politics	Economics	Art & Entertainment	Sports	Science & Health	Technology	Unclassified
Arab Countries	6.2	5.6	16.2	7.2	11	3.9	49.9
Argentina	3.1	23.1	15.9	8.9	7.3	8.9	32.8
Australia	8.2	7	21.4	9.8	12.6	8.6	32.5
Brazil	4.7	10.4	15.9	14.8	11.6	7.1	35.6
Canada	3.5	9.6	21.9	11.5	13.1	9	31.3
China	5.4	10.3	15.4	6.2	10.2	10.9	41.7
Colombia	7.4	6.9	13.8	13.1	11.6	8.2	39
France	5.2	10.7	13.1	10.2	6.9	16.2	37.7
Germany	3.6	28	23.9	6.8	6.9	5.4	25.4
India	6.6	5.7	12.3	10.1	8.9	9.8	46.5
Indonesia	5.5	4.7	13.4	10.8	8.4	11.6	45.6
Iran	9.4	8.9	12.7	10	10	15.4	33.7
Italy	4.3	15.7	25.2	9.9	10.1	9.4	25.3
Japan	4.3	5.4	18.8	11.9	14.4	10.2	34.9
Korea	4	4.5	21.1	11.2	16.6	6.4	36
Philippines	5.5	5.8	14.7	6.8	7.3	7	52.9
Russia	6.4	13.1	19.2	10	11.7	7.9	31.7
Spain	5.3	14.1	15.9	10.6	11.7	9.3	33
Thailand	4.6	8.5	13.4	11.2	9.3	6.3	46.7
Turkey	7.2	5.7	16.9	12.8	14.7	8	34.6
UK	4.4	7.9	21.3	13.8	13.4	9.1	30
USA	2.6	4.9	27.7	13	16.6	6.9	28.2
Average	5.3	9.8	17.7	10.5	11.1	8.9	36.6

C. Detecting Country of Twitter Users

In our cross-cultural framework, we need to separate data by their origin country. Unfortunately, it is reported by multiple studies [35]–[37] that less than 5% of all tweets have geo-location tags, which is far from an ideal point for research studies.

TABLE V
CULTURAL DISTANCE BETWEEN THE COUNTRIES BASED ON THE RESULTS OF THIS STUDY

	Arab Countries	Argentina	Australia	Brazil	Canada	China	Colombia	France	Germany	India	Indonesia	Iran	Italy	Japan	Korea	Philippines	Russia	Spain	Thailand	Turkey	UK	USA
Arab Countries	0	25.5	19.1	17.3	21.2	11.9	13.4	19.1	34.4	8.6	10.3	20.9	28.7	17.6	16.6	6	20.4	19.9	7.3	17.3	22.4	26.2
Argentina	25.5	0	18.6	15.2	16.2	16.4	19	15.7	12.6	22.8	22.9	17.4	14.5	19.8	22	26.8	12.1	10.4	20.7	19.9	18.3	24.5
Australia	19.1	18.6	0	9.6	5.9	12.8	10.6	14.5	23.9	17.3	16.6	11.6	12.8	6.4	7.7	22.6	6.8	9.6	17.3	6.4	6.2	11
Brazil	17.3	15.2	9.6	0	8.6	11.4	6.3	11.8	23.8	13.9	13.6	11.5	15.9	7.8	10.1	20.1	7.7	6.6	12.3	6.7	8.7	16
Canada	21.2	16.2	5.9	8.6	0	14	12.4	14.7	21.2	19.2	18.3	13.4	9.8	6.6	8.3	24.4	5.8	8.1	18.2	8.4	3.3	9.2
China	11.9	16.4	12.8	11.4	14	0	9.1	8.8	26.3	8.6	8.7	11.2	20.2	11.6	13.6	13	12.1	10.7	8.9	12.3	15.8	21.7
Colombia	13.4	19	10.6	6.3	12.4	9.1	0	10.8	28.5	8.9	8.9	10.1	20.5	8.3	10.6	16	11.4	10.2	9.1	6.4	12.4	19
France	19.1	15.7	14.5	11.8	14.7	8.8	10.8	0	26.4	12.3	11.1	6.9	19.5	12.8	17.3	18.8	13.1	10.6	13.9	13.7	15.5	23
Germany	34.4	12.6	23.9	23.8	21.2	26.3	28.5	26.4	0	33.4	33.4	26.7	13.8	27	28	36.6	18.2	19.2	31.1	27.3	23.1	26.2
India	8.6	22.8	17.3	13.9	19.2	8.6	8.9	12.3	33.4	0	2.9	14.7	26.8	14.8	16.4	8.2	18.2	16.6	5.2	14.5	19.9	25.7
Indonesia	10.3	22.9	16.6	13.6	18.3	8.7	8.9	11.1	33.4	2.9	0	13.9	26.1	13.6	15.7	9.7	17.9	16.4	6.8	14	18.9	24.6
Iran	20.9	17.4	11.6	11.5	13.4	11.2	10.1	6.9	26.7	14.7	13.9	0	18.3	11.3	15.8	22	11.5	9.8	16.7	10.9	13.5	21
Italy	28.7	14.5	12.8	15.9	9.8	20.2	20.5	19.5	13.8	26.8	26.1	18.3	0	16.1	17.6	31.5	9.7	12.4	25.7	17.1	11.1	13.8
Japan	17.6	19.8	6.4	7.8	6.6	11.6	8.3	12.8	27	14.8	13.6	11.3	16.1	0	5.2	20.7	9.5	9.9	14.9	4.3	6.5	11.9
Korea	16.6	22	7.7	10.1	8.3	13.6	10.6	17.3	28	16.4	15.7	15.8	17.6	5.2	0	20.9	11.3	12.8	15.6	6.4	8.5	10.5
Philippines	6	26.8	22.6	20.1	24.4	13	16	18.8	36.6	8.2	9.7	22	31.5	20.7	20.9	0	23.5	22.5	8.5	20.9	25.8	30.2
Russia	20.4	12.1	6.8	7.7	5.8	12.1	11.4	13.1	18.2	18.2	17.9	11.5	9.7	9.5	11.3	23.5	0	4.1	17	9.2	7.5	14.1
Spain	19.9	10.4	9.6	6.6	8.1	10.7	10.2	10.6	19.2	16.6	16.4	9.8	12.4	9.9	12.8	22.5	4.1	0	15.5	9.7	9.6	17
Thailand	7.3	20.7	17.3	12.3	18.2	8.9	9.1	13.9	31.1	5.2	6.8	16.7	25.7	14.9	15.6	8.5	17	15.5	0	14.5	19.3	24.8
Turkey	17.3	19.9	6.4	6.7	8.4	12.3	6.4	13.7	27.3	14.5	14	10.9	17.1	4.3	6.4	20.9	9.2	9.7	14.5	0	7.6	13.5
UK	22.4	18.3	6.2	8.7	3.3	15.8	12.4	15.5	23.1	19.9	18.9	13.5	11.1	6.5	8.5	25.8	7.5	9.6	19.3	7.6	0	8.4
USA	26.2	24.5	11	16	9.2	21.7	19	23	26.2	25.7	24.6	21	13.8	11.9	10.5	30.2	14.1	17	24.8	13.5	8.4	0

We use a method proposed by Van der Veen *et al.* [35] to determine the country of users that did not specify a valid location in their profile. This method that achieved 82% accuracy consists of a classification model that is trained by user time zone, language, and location string. Authors identified limited information of user profiles and shared properties between countries, such as shared time zone and language, as the root causes of errors in the country identification process.

D. Translation of Input Texts

This study faces the problem of multilingual inputs, which causes problems in the process of text classification. To overcome this problem, we propose the translation of all input texts from their original language to a single target language using online translation services. Based on our observations on the results of Google Translation Service, in case the input text was already written in the chosen language, the translation will fix some misspellings of the input text.

We propose to use the English language as the target language, since it has been shown by a benchmark [38] that online translation services perform with the highest accuracy when the target language is English.

E. Tweet Classification

In Section II-C, we explained the main challenges of our text classification process. Due to the informality, sparsity, multilingualism, and the broad context of our data set, most text classification techniques become unusable in this paper.

Due to the informality, sparsity, and broad context of the data set, most of the text classification techniques are incompatible with the needs of this study. Since Naïve Bayes classifiers are a common approach when it is difficult to extract meaningful relations between the words of the input text [30], we used a variation of it as the core for the classification process of our framework.

In this classification algorithm, if t is a tweet and C is the set of categories, the main class of t is calculated as

$$\text{MainClass}(t, C) = \underset{c \in C}{\text{argmax}}(p(t|c)) \tag{1}$$

$$p(t|c) = \frac{\left(\sum_{w \in t} f_i\right)!}{\prod_{w \in t} f_i!} \prod_{w \in t} p(w|t)^{f_i} \tag{2}$$

$$p(w|c) = \frac{\text{TFIDF}(w, c) + 1}{\left(\sum_{v \in C} \text{TFIDF}(v, c)\right) + |\{v \in C\}|} \tag{3}$$

where f_i is the frequency count of word i in tweet t.

Since not every tweet has news orientation, we specified a minimum threshold to separate news-oriented tweets from the rest. We used heuristics to find this threshold by examining its effect on the overall precision of the classification algorithm. Based on our assessment, we considered a tweet as *Unclassified*, if its highest value of probability in 1 is less than 40%.

As we mentioned earlier, we use TF-IDF[1] method to quantify the importance of a word to a category. The method is a combination of two values: the term frequency and the inverse document frequency. The first statistic equals the frequency of the word in the training documents of a particular category, while the second statistic diminishes the weight of common words that occur frequently in all categories (such as conjunctions). To be specific, IDF computes logarithmically scaled inverse fraction of the classes that contain the word. Therefore, common words that appear in most classes will

[1] Short for term frequency-inverse document frequency [40].

TABLE VI
ACCURACY OF THE TEXT CLASSIFICATION ALGORITHM

Country	Accuracy (percent)
Arab Countries	87
Argentina	90
Australia	92
Brazil	91
Canada	93
China	87
Colombia	88
France	83
Germany	91
India	90
Indonesia	83
Iran	84
Italy	92
Japan	79
Korea South	86
Philippines	93
Russia	91
Spain	85
Thailand	77
Turkey	82
United Kingdom	94
USA	96
Average	**87.9**

have a low impact on the process and words that are most likely to be used in one class gain a considerable weight in that particular class [39].

As a side note and based on our observation of users tweeting behavior, users apply abbreviation on the common words of their tweet and not on its keywords. Since common words have little or no effect in most of the text classification algorithms, especially in Naïve Bayes, these abbreviations have a low impact on the final output of the classification. We also apply TF-IDF algorithm to further decrease the effect of common words on the output of the classification and, finally, the results of distribution.

F. Cultural Distance

To discuss cross-cultural similarities and differences, we propose a method to quantify cultural closeness of a country to another based on the results of the previous step. This is an attempt to automatically interpret the results of tweet distribution and to find cultural similarities and differences. To this end, we consider overall distribution vector of a country as its location in a 7-D space, and the Euclidean distance between any two countries as their cultural distance.

Therefore, if $A = (a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ is the distribution results of a country (where a_1–a_6 are the percentage of tweets in the selected categories and a_7 is the percentage of unclassified tweets) and $B = (b_1, b_2, b_3, b_4, b_5, b_6, b_7)$ is the distribution results of a second country, the cultural distance between the two countries of A and B is computed as

$$\text{CulD}(A, B) = \sqrt{\sum_{i=1}^{|C|+1} (a_i - b_i)^2} \tag{4}$$

where C is the set of categories.

TABLE VII
HOFSTEDE'S CULTURAL SCORES FOR THE SELECTED COUNTRIES (VALUES RANGE: 0–120)

	Power distance	Individualism	Masculinity	Uncertainty avoidance	Long-term orientation	Indulgence
Arab Countries	80	38	53	68	23	34
Argentina	49	46	56	86	20	62
Australia	38	90	61	51	21	71
Brazil	69	38	49	76	44	59
Canada	39	80	52	48	36	68
China	80	20	66	30	87	24
Colombia	67	13	64	80	13	83
France	68	71	43	86	63	48
Germany	35	67	66	65	83	40
India	77	48	56	40	51	26
Indonesia	78	14	46	48	62	38
Iran	58	41	43	59	14	40
Italy	50	76	70	75	61	30
Japan	54	46	95	92	88	42
Korea	60	18	39	85	100	29
Philippines	94	32	64	44	27	42
Russia	93	39	36	95	81	20
Spain	57	51	42	86	48	44
Thailand	64	20	34	64	32	45
Turkey	66	37	45	85	46	49
UK	35	89	66	35	51	69
USA	40	91	62	46	26	68
Average	**61.4**	**48.4**	**54.9**	**65.6**	**49**	**46.9**

IV. EXPERIMENT RESULTS

Using Twitter's API, we collected 17 065 069 tweets from November to December 2017. Our data set consists of all available data about tweets and their authors, including tweet identifier, text, language, time, time zone, username, and location.

We trained a model based on [35] to separate tweets of a country by using related user information, such as language, location, and time zone. From all of the gathered tweets, our model mapped 9 197 071 tweets to one of the 22 selected countries. The dictionary was unable to identify the origin country of users for a portion of tweets for several reasons: First, we only examined 22 countries. Second, authors of many tweets have not specified a location or time zone in their profiles. Finally, our model had difficulty in separating tweets of some neighbor countries: for example, many users from America time zones tweeted in English without specifying any location or country.

Table III represents the statistics on the total number of tweets (including retweets), the number of original tweets and percentage of original tweets per country. Average for the original tweets is 53.9%, meaning that more than 46% of our collected tweets in these countries were retweets.

A. Results of Distribution

We translated our data set into English and applied the classification process on the translated input. Table IV represents the distribution of tweets in the selected $6 + 1$ categories for each of the 22 countries.

TABLE VIII
CULTURAL DISTANCE BETWEEN THE COUNTRIES BASED ON HOFSTEDE'S CULTURAL SCORES (TABLE VII)

	Arab Countries	**Argentina**	**Australia**	**Brazil**	**Canada**	**China**	**Colombia**	**France**	**Germany**	**India**	**Indonesia**	**Iran**	**Italy**	**Japan**	**Korea**	**Philippines**	**Russia**	**Spain**	**Thailand**	**Turkey**	**UK**	**USA**
Arab Countries	0	46.3	78.9	35.6	72.1	78.7	59.7	59	81.7	41.7	50.6	28.3	64.7	85.8	85.1	31.8	69.2	42.9	34	35.9	89.1	78.4
Argentina	46.3	0	58.3	34.2	54.8	104	44.9	56.6	74.4	71.5	75.6	38.9	62.7	80.9	92.6	67	89.1	36.9	47.7	35.9	75.8	61.7
Australia	78.9	58.3	0	71.4	20.5	117	88.4	71.8	74.4	79.8	102	64.9	66.1	101	123	86.4	120	70.3	85.3	78.1	34.5	7.9
Brazil	35.6	34.2	71.4	0	60	77.5	48.8	41.7	65.4	51.5	47.3	41.5	58.5	70.1	68.5	49.7	63.7	26.8	32.6	14.5	76.7	71.6
Canada	72.1	54.8	20.5	60	0	102	84.8	60.2	60.4	67.4	86.7	58	57.7	92.5	106	79.1	107	58.8	73.2	66.8	26.3	18.2
China	78.7	104	117	77.5	102	0	108	87	75.9	48.3	40.1	89.6	82.4	79.8	66	66.9	75.5	84.7	77.4	79.7	101	112
Colombia	59.7	44.9	88.4	48.8	84.8	108	0	87.3	104	87.5	76.9	59.7	97.5	98.3	105	65.4	104	69.4	54.8	56.3	102	91.4
France	59	56.6	71.8	41.7	60.2	87	87.3	0	50.1	59.3	70	65.4	39.1	64.8	67.5	75.7	53.6	28.1	64.8	38.6	71.9	70.6
Germany	81.7	74.4	74.4	65.4	60.4	75.9	104	50.1	0	63.9	76.1	81	31.5	49	67.6	90.7	79.8	54.9	81.9	64.6	56.9	70.8
India	41.7	71.5	79.8	51.5	67.4	48.3	87.5	59.3	63.9	0	39.7	50.3	55.3	79.8	77	37.7	68.8	55.1	52.3	54.3	73.8	75.4
Indonesia	50.6	75.6	102	47.3	86.7	40.1	76.9	70	76.1	39.7	0	60	77.4	81.4	57.2	46.1	62.1	59.2	40	49.4	95.7	99.3
Iran	28.3	38.9	64.9	41.5	58	89.6	59.7	65.4	81	50.3	60	0	68.4	96.7	93.9	47.3	87.1	44.7	30.6	43.1	78.7	65.3
Italy	64.7	62.7	66.1	58.5	57.7	82.4	97.5	39.1	31.5	55.3	77.4	68.4	0	51.7	77.5	78.6	72.6	44.3	76.6	56	60.8	63.1
Japan	85.8	80.9	101	70.1	92.5	79.8	98.3	64.8	49	79.8	81.4	96.7	51.7	0	65.6	93.4	74.7	67.1	91.9	68	91.8	100
Korea	85.1	92.6	123	68.5	106	66	105	67.5	67.6	77	57.2	93.9	77.5	65.6	0	95.3	45.6	63.6	73.4	61.5	113	121
Philippines	31.8	67	86.4	49.7	79.1	66.9	65.4	75.7	90.7	37.7	46.1	47.3	78.6	93.4	95.3	0	82.6	66.2	48.7	56.8	90.2	84.2
Russia	69.2	89.1	120	63.7	107	75.5	104	53.6	79.8	68.8	62.1	87.1	72.6	74.7	45.6	82.6	0	57.1	72.6	55.2	117	118
Spain	42.9	36.9	70.3	26.8	58.8	84.7	69.4	28.1	54.9	55.1	59.2	44.7	44.3	67.1	63.6	66.2	57.1	0	42.6	17.9	76.1	70.5
Thailand	34	47.7	85.3	32.6	73.2	77.4	54.8	64.8	81.9	52.3	40	30.6	76.6	91.9	73.4	48.7	72.6	42.6	0	32.6	91.8	85.4
Turkey	35.9	35.9	78.1	14.5	66.8	79.7	56.3	38.6	64.6	54.3	49.4	43.1	56	68	61.5	56.8	55.2	17.9	32.6	0	84	78.5
UK	89.1	75.8	34.5	76.7	26.3	101	102	71.9	56.9	73.8	95.7	78.7	60.8	91.8	113	90.2	117	76.1	91.8	84	0	28.5
USA	78.4	61.7	7.9	71.6	18.2	112	91.4	70.6	70.8	75.4	99.3	65.3	63.1	100	121	84.2	118	70.5	85.4	78.5	28.5	0

Category of Art and Entertainment contains the highest amount of tweets among the news categories (close to 18% of all tweets). USA with more than 27% and Germany with 23.9% have the highest ratios in this category. Germany, with a ratio of 28%, has also the highest percentage of tweets in the Economics category. Politics category contains the smallest portion of tweets with less than 5% of all tweets.

The last column of Table IV is dedicated to tweets that our classifier was unable to classify as a news-oriented tweet. We associate this column with three types of tweets:

1) Tweets that did not have any news orientation and no class reached the minimum threshold of the classification algorithm (Section III-D). Therefore, the classifier correctly labeled them as *Unclassified*.
2) Tweets that did not include enough information to be processed in the classification algorithm. Some examples of this type are tweets with less than three words, or tweets of pictures and videos without any accompanying text.
3) Tweets with news orientation that our classifier was unable to determine the category, meaning that none of the categories in the text classification algorithm reached the minimum threshold of acceptance. Thus, the classifier was mistaken in labeling them as *Unclassified*.

B. Results of Cultural Distance

By considering the computed distributions of tweets over the six categories (from Table IV) as a country's location in a 7-D space (six news categories and one column for unclassified tweets), we calculated cultural distance between all of the selected countries.

The results of this calculation are displayed in Table V. The lowest values of cultural distance are for the pairs of Arab Countries–Philippines and Japan–U.K., and the highest ones are for Germany–Philippines and Germany–Arab Countries.

V. EVALUATION

We evaluate our framework using two separate approaches: First, we examine the accuracy of the text classification process, wherein we manually classified a portion of the input data set, separately for all of the selected countries. Second, we evaluate the results of cultural distance by comparing the results of our study to the cultural distances computed using the Hofstede's latest national scores. The first evaluation is merely to seek the accuracy of the classification method, while the second one is for the overall speculation of the final results and the whole framework.

A. Evaluation of the Text Classification Algorithm

In order to evaluate the accuracy of the classification algorithm, we manually classified more than 2500 tweets in total (more than 100 tweets for each country).

Table VI represents the precision of the classifier for each country. In average, our classifier determined the correct category of a tweet with close to 70%. This is a fair and acceptable precision if we consider the combined challenges of the classification process of this research (short text, informal text, multilingual inputs, and no context).

B. Evaluation of Cultural Distances

In order to evaluate the final results of our framework, we compared cultural distances computed based on the results of this study with cultural distances computed based on the Hofstede's cultural scores.

Therefore, we collected the latest scores of Hofstede cultural dimensions for the selected nations of this study and

TABLE IX
RESULTS OF CORRELATION ANALYSIS BETWEEN CULTURAL DISTANCES COMPUTED BASED ON THE RESULTS OF THE TWO STUDIES

Country	Correlation Coefficient
Arab Countries	0.581
Argentina	0.412
Australia	0.281
Brazil	0.318
Canada	0.37
China	0.544
Colombia	0.443
France	0.292
Germany	0.711
India	0.573
Indonesia	0.701
Iran	0.216
Italy	0.581
Japan	0.157
Korea South	0.175
Philippines	0.797
Russia	0.164
Spain	0.369
Thailand	0.622
Turkey	0.13
United Kingdom	0.417
USA	0.47
Average	**0.424**

TABLE X
RESULTS OF CORRELATION ANALYSIS BETWEEN THE DISTRIBUTION RESULTS OF THIS STUDY AND HOFSTEDE'S CULTURAL SCORES (COEFFICIENTS)

	Power distance	Individualism	Masculinity	Uncertainty avoidance	Long-term orientation	Indulgence
Politics	0.357	-0.257	-0.223	-0.014	-0.323	-0.076
Economics	-0.327	0.216	0.045	0.252	0.145	-0.059
Art & Entertainment	**-0.642**	**0.664**	0.35	-0.049	0.217	0.177
Sports	-0.291	0.135	-0.085	0.204	-0.136	**0.536**
Science	-0.303	0.165	0.084	0.104	0.111	0.236
Technology	0.027	0.082	-0.053	-0.03	0.049	-0.06

used them as the second group of indicators for a nation's culture. Table VII shows the cultural scores of Hofstede for the selected countries. Using this new indicator (Table VII), we calculated cultural distances between the selected nations (available in Table VIII). The lowest values of cultural distance are for the pairs of Arab Countries-Philippines and USA-U.K., and the highest ones are for Japan-Philippines and Japan-USA. It should be noted that these new distances (values of Table VIII) are relatively higher in quantity than the values in Table V, since Hofstede scores range from 0 to 120.

We used correlation analysis to examine the existence of relationships between the two groups of cultural distances. We did this separately for each country by inserting distance values of a country in a two-column table and analyzing correlation of the values of the two columns. The result of correlation analysis is represented in Table IX. Interestingly, it shows a positive correlation for all of the 22 countries. Correlation results for four countries (Arab Countries, Germany, Indonesia, and Philippines) can be considered as strong. Finally, it should be noted that as normalization has no effect in correlation analysis, we did not normalize or alter Hofstede's scores in any way.

TABLE XI
RESULTS OF CORRELATION ANALYSIS BETWEEN THE SELECTED NEWS CATEGORIES BASED ON THE DISTRIBUTION RESULTS OF THIS STUDY (COEFFICIENTS)

	Politics	Economics	Art & Entertainment	Sports	Science & Health	Technology
Politics	1	-0.340	**-0.509**	-0.0657	-0.083	0.323
Economics	-0.340	1	0.214	-0.367	**-0.502**	-0.066
Art & Entertainment	**-0.509**	0.214	1	0.118	**0.507**	-0.386
Sports	-0.065	-0.367	0.118	1	**0.584**	0.040
Science	-0.083	**-0.502**	**0.507**	**0.584**	1	-0.269
Technology	0.323	-0.066	-0.386	0.040	-0.269	1

Evaluation results clearly show that there is an absolute connection between the cultural distances computed based on cultural indicators of this study and the cultural distances computed based on Hofstede's cultural scores. This remark can be interpreted as a proof of concept or legitimacy of the proposed framework.

C. Relationships Between Cultural Indicators

In this final step, we will examine possible connections between cultural indicators. To this aim, first, we correlated cultural scores of our framework (tweets distributions in Table IV) with scores of Hofstede's cultural study (Table VII). Correlation coefficients for all possible connections are presented in Table X. In addition, we analyzed connections between a country's tendencies toward a news category with other news categories. For this part, we correlated raw distribution results of this study with each other (results available in Table XI).

These correlation analyses found four strong connections. These connections are summarized in the following.

1) The individualism index of Hofstede was positively correlated with the tendency of users toward Art-oriented tweets (coefficient of 0.72), which supports or is supported by theories about the connection of artistic behavior and individualism [41]–[43].
2) Tendency of users toward Art and Entertainment was negatively correlated with the power distance index of Hofstede (coefficient of -0.62), thus countries with higher power distance (as defined by Hofstede) would have less tendency to tweet about Art.
3) The scores for science category is in negative correlation with sports category (coefficient of -0.661), meaning that countries with a higher percentage of tweets in science and health category are relatively less likely to tweet about sports.

4) The tendency of a country toward Art and Entertainment is in negative correlation with the tendency toward political news (coefficient of -0.55).

VI. Conclusion

We identified major issues in traditional methods of data collection in cross-cultural studies, which include the need for access to people from many nations, a limited number of samples, negative impacts of translation, positive self-enhancement illusions, and some unreported problems. These issues are either causing difficulty to perform a cross-cultural study or have negative impacts on the validity of the final results.

In this paper, we presented an automated framework to measure the cultural distance between several countries using the information extracted from social networks. In its technical form, the framework has several stages. First, it trains a classification model using previously classified news articles and tweets. Second, it collects random tweets to form the input data set and separates them by nationality of the user. Then, it translates the tweets and the collected training data set to a single language and applies the classification process on all of input tweets. Finally, after aggregating the result of the classification process for each nation, it uses the distribution of tweets over categories as the cultural indicators or factors of a country and, thus, computes the cultural distances using these aggregated values of distribution. By exploring major news broadcasting websites, we picked six news categories to be used as classes in the classification algorithm. To determine the possible nationality of a twitter user, we trained a model to map set of users' available data (such as language, time zone, and location) to the most-likely country. We applied Multinomial Naïve Bayes classification algorithm to classify all input data and extract the distribution of tweets over the six categories. We also proposed an online resource to automatically collect training data set. To overcome the challenge of multilingualism, we preferred to translate all data (training and input) to a single language.

By applying the proposed framework to a sample of 17 M tweets, we computed cultural distance between 22 countries. Our model mapped more than 51% of the tweets to one of the selected countries. Our analysis showed that more than 46% of tweets were retweets and it reaches as high as 88% in Thailand. The text classification algorithm maintained a relatively high accuracy for most of the countries (with an average close to 88%) considering the classification challenges of this study (short, informal, multilingual, and the broad context of tweets).

Correlation analysis found four strong connections among the cultural indicators of this study and Hofstede's cultural scores: Tendency of users toward Art and Entertainment news is correlated with the power distance (coefficient of -0.62) and individualism index of Hofstede (coefficient of 0.72). This tendency is also in negative correlation with the tendency toward political news (coefficient of -0.55). Tendency of users toward scientific news is in negative correlation with sports category (coefficient of -0.661).

We evaluated the final results of our framework by correlating cultural distances computed based on the results of this study and cultural distances based on Hofstede's cultural scores. We did this separately for each of the selected countries and interestingly, all of them showed a positive correlation (with an average of 0.46). This remark can be interpreted as a proof of concept or legitimacy of our proposed framework in mining social networks in conducting a cross-cultural research. While in this paper, we only used an automated model to discuss the cultural differences and similarities of our results, it is needless to say that other methods of interpretation can be used to discuss results of distribution and cultural distances.

Finally, we like to explore different aspects of our method of data collection in regards to the above-mentioned issues. By collecting data from social networks, there was no requirement to find participants from different nations using time- and energy-consuming methods. Our main difficulty in accessing people from different nations was the problem of detecting country of a user, which we handled using a classification model. Our number of samples (225 K per country, in average) cannot be compared to the previous cross-cultural studies (less than 1200 samples per country). While we translated all input data to a target language, this translation was merely on the content and not on the questions. Furthermore, since the users of social networks are acting in their normal state, we have no reason to believe that positive self-enhancement illusion has an impact on social networks.

The proposed framework can be used as a technical or abstract framework to perform cross-cultural studies by mining social networks data. Future cross-cultural works need to design and use mining/analysis models to extract cultural features from social networks with respect to their research questions. In addition, computer scientists can improve precision of our models in detecting country of a user or classification of a tweet.

References

[1] S. Papayiannis and X. Anastassiou-Hadjicharalambous, "Cross-cultural studies," in *Encyclopedia of Child Behavior and Development*, S. Goldstein, J. A. Naglieri, Eds. Boston, MA, USA: Springer, 2011, pp. 438–440.

[2] S. H. Schwartz and A. Bardi, "Value hierarchies across cultures: Taking a similarities perspective," *J. Cross-Cultural Psychol.*, vol. 32, no. 3, pp. 268–290, 2001.

[3] A. Terracciano *et al.*, "National character does not reflect mean personality trait levels in 49 cultures," *Science*, vol. 310, no. 5745, pp. 96–100, 2005.

[4] G. Hofstede, "Cultural dimensions in management and planning," *Asia Pacific J. Manage.*, vol. 1, no. 2, pp. 81–99, 1984.

[5] P. B. Smith, S. Dugan, and F. Trompenaars, "National culture and the values of organizational employees: A dimensional analysis across 43 nations," *J. Cross-Cultural Psychol.*, vol. 27, no. 2, pp. 231–264, 1996.

[6] L. Sechrest, T. L. Fay, and S. H. Zaidi, "Problems of translation in cross-cultural research," *J. Cross-Cultural Psychol.*, vol. 3, no. 1, pp. 41–56, 1972.

[7] A. H. de Mendoza, "The problem of translation in cross-cultural research on emotion concepts (Commentary on Choi & Han)," *Int. J. Dialogical Sci.*, vol. 3, no. 1, pp. 241–248, 2008.

[8] D. P. Schmitt, A. Realo, M. Voracek, and J. Allik, "Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures," *J. Personality Social Psychol.*, vol. 94, no. 1, p. 168, 2008.

[9] R. R. McCrae and A. Terracciano, "Personality profiles of cultures: Aggregate personality traits," *J. Personality Social Psychol.*, vol. 89, no. 3, p. 407, 2005.

[10] D. P. Schmitt, J. Allik, R. R. McCrae, and V. Benet-Martínez, "The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations," *J. Cross-Cultural Psychol.*, vol. 38, no. 2, pp. 173–212, 2007.

[11] S. E. Taylor and J. D. Brown, "Illusion and well-being: A social psychological perspective on mental health," *Psychol. Bull.*, vol. 103, no. 2, p. 193, 1988.

[12] O. P. John and R. W. Robins, "Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism," *J. Personality Social Psychol.*, vol. 66, no. 1, p. 206, 1994.

[13] E. Callahan, "Cultural similarities and differences in the design of University Web sites," *J. Comput.-Mediated Commun.*, vol. 11, no. 1, pp. 239–273, 2005.

[14] J. Park, Y. M. Baek, and M. Cha, "Cross-cultural comparison of nonverbal cues in emoticons on Twitter: Evidence from big data analysis," *J. Commun.*, vol. 64, pp. 333–354, Apr. 2014.

[15] P. B. Smith, F. Trompenaars, and S. Dugan, "The rotter locus of control scale in 43 countries: A test of cultural relativity," *Int. J. Psychol.*, vol. 30, no. 3, pp. 377–400, 1995.

[16] A. Merritt, "Culture in the Cockpit: Do Hofstede's dimensions replicate?" *J. Cross-Cultural Psychol.*, vol. 31, no. 3, pp. 283–301, 2000.

[17] U. Schimmack, S. Oishi, and E. Diener, "Cultural influences on the relation between pleasant emotions and unpleasant emotions: Asian dialectic philosophies or individualism-collectivism?" *Cognition Emotion*, vol. 16, no. 6, pp. 705–719, 2002.

[18] N. Struch, S. H. Schwartz, and W. A. Van Der Kloot, "Meanings of basic values for women and men: A cross-cultural analysis," *Personality Social Psychol. Bull.*, vol. 28, no. 1, pp. 16–28, 2002.

[19] M. H. Bond *et al.*, "Culture-level dimensions of social axioms and their correlates across 41 cultures," *J. Cross-Cultural Psychol.*, vol. 35, no. 5, pp. 548–570, 2004.

[20] R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, and V. Gupta, *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Newbury Park, CA, USA: Sage, 2004.

[21] P. Kuppens, E. Ceulemans, M. E. Timmerman, E. Diener, and C. H. U. Kim-Prieto, "Universal intracultural and intercultural dimensions of the recalled frequency of emotional experience," *J. Cross-Cultural Psychol.*, vol. 37, no. 5, pp. 491–515, 2006.

[22] C. Welzel, "How selfish are self-expression values? A civicness test," *J. Cross-Cultural Psychol.*, vol. 41, no. 2, pp. 152–174, 2010.

[23] R. Basak, S. Sural, N. Ganguly, and S. K. Ghosh, "Online public shaming on twitter: Detection, analysis, and mitigation," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 2, pp. 208–220, Apr. 2019.

[24] Y. Liu and S. Xu, "Detecting rumors through modeling information propagation networks in a social media environment," *IEEE Trans. Comput. Social Syst.*, vol. 3, no. 2, pp. 46–62, Jun. 2016.

[25] J. Cole, M. Ghafurian, and D. Reitter, "Word adoption in online communities," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 1, pp. 178–188, Feb. 2019.

[26] F. Liu, Y. Liu, and F. Weng, "Why is 'SXSW' trending?: Exploring multiple text sources for Twitter topic summarization," in *Proc. Workshop Lang. Social Media*, Association for Computational Linguistics, 2011, pp. 66–75.

[27] B. O'Connor, M. Krieger, and D. Ahn, "Tweetmotif: Exploratory search and topic summarization for twitter," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 384–385.

[28] R. Zhang, W. Li, D. Gao, and Y. Ouyang, "Automatic Twitter topic summarization with speech acts," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 649–658, Mar. 2013.

[29] D. L. Lasorsa, S. C. Lewis, and A. E. Holton, "Normalizing Twitter: Journalism practice in an emerging communication space," *Journalism Stud.*, vol. 13, no. 1, pp. 19–36, 2012.

[30] Y. Li, A. Tripathi, and A. Srinivasan, "Challenges in short text classification: The case of online auction disclosure," in *Proc. MCIS*, 2016, p. 18.

[31] J. Szymański, "Self–organizing map representation for clustering Wikipedia search results," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Berlin, Germany: Springer, 2011, pp. 140–149.

[32] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proc. 3rd Annu. Symp. Document Anal. Inf. Retr.*, Ann Arbor, MI, USA, 1994, vol. 48113, no. 2, pp. 161–175.

[33] W. Cavnar, "Using an n-gram-based document representation with a vector processing retrieval model," in *Proc. NIST*, 1995, p. 269.

[34] F. Hu, Z. Shao, and T. Ruan, "Self-supervised chinese ontology learning from online encyclopedias," *Sci. World J.*, vol. 2014, Mar. 2014, Art. no. 848631.

[35] H. Van der Veen, D. Hiemstra, T. van den Broek, M. Ehrenhard, and A. Need, "Determine the user country of a tweet," 2015, *arXiv:1508.02483*. [Online]. Available: https://arxiv.org/abs/1508.02483

[36] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of Twitter users," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 47, 2014.

[37] M. Marciniec. (2017). *Observing World Tweeting Tendencies in Real-Time—Part 2*. [Online]. Available: https://codete.com/blog/observing-world-tweeting-tendencies-in-real-time-part-2/

[38] Intento. (Jul. 2017). *Intento Machine Translation Benchmark*. [Online]. Available: https://www.slideshare.net/KonstantinSavenkov/intento-machine-translation-benchmark-july-2017

[39] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[40] J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Datasets*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[41] J. Elkins, *Why Art Cannot Be Taught: A Handbook for Art Students*. Champaign, IL, USA: Univ. Illinois Press, 2001.

[42] O. Wilde, *The Soul of Man Under Socialism and Selected Critical Prose*. London, U.K.: Penguin, 2001.

[43] R. Wittkower, "Individualism in art and artists: A renaissance problem," *J. History Ideas*, vol. 22, no. 3, pp. 291–302, 1961.