Torun Business Review 16(2) 2017 63-73

CULTURAL CONTEXT IN PROCESS OF MINING DATA FROM SOCIAL MEDIA – RECOMMENDATIONS BASED ON LITERATURE REVIEW

JOANNA MICHALAK[a], PATRYCJA GULAK-LIPKA[a],
[a] *Nicolaus Copernicus University in Torun, Poland*

ABSTRACT

Social media is nothing else than a modern communication channel that carries many advantages, such as reach or range. Social media has such a big power of its reach that a single post, tweet, or "broad" becomes a global issue. With globalization, we have seen an increase in usage of social media everywhere. This means that communication happens across the borders of different countries, continents or even cultures. It is a desirable effect, however, worldwide social media users differ with respect to their culture and data shows that significant differences exist in the way people in the world use social media. However, in order to be well prepared to dig in the social media, a question should be asked whether the cultural context affects the activity of users. If it does, it is appropriate to prepare data filters to include some specific criteria. In the first part, the authors apply the Cross - Industry Standard Process for Data Mining (CRISP-DM) in social media data to specify the process of data analysis. The second part focuses on the recommendations about cultural context in social media mining.

ARTICLE INFO

Available online 23 May 2017

Keywords:
social media,
Twitter,
CRISP-DM,
cultural context,
Web 2.0.

JEL: C80, M31.

Doi: 10.19197/tbr.v16i2.109

INTRODUCTION

The process of analyzing social media data is extremely complex. The correct specification of the purpose of the analysis is important – social media reflect a huge data stream and it is nearly impossible for anyone to look through it, so well-chosen methodology is very important. The basic measure for mining such data is transforming it into a struc-

ture that can be understood. Extracted knowledge is the knowledge about users' activity concerning events, organizations, brands etc. However, in order to be well prepared to dig into social media, the question should be asked whether the cultural context affects the activity of users. If it does, it is appropriate to prepare data filters to include some specific criteria. In the first part, the authors apply the Cross-Industry Standard Process for Data Mining (CRISP-DM) to social media data to specify the process of data analysis and difficulties associated with it. The second part focuses on cultural context and its localization of in the mining of social media process. Finally, brief recommendations are made. It must be noted that authors do not specify business questions but by looking at general problem, they are trying to form some recommendations that can be the starting point for further discussion.

CROSS – INDUSTRY STANDARD PROCESS IN CASE OF TWITTER DATA MINING

Currently, various authors widely debate the approach to mining datasets containing data from social media. According to (Larose, (2006)) and (Gartner Group, www.gartner.com/technology/home.jsp, 31.03. 2017): data mining is a process of discovering important and new patterns/trends when searching through large amounts of data stored in databases using statistical and mathematical methods. Data mining research without prior preparation of data can lead to erroneous conclusions or wrong approach can be applied to the dataset. This has results in incurring costs for the organization. According to (Larose, (2006)) improper analysis is worse than lack of an analysis.

Cross-Industry Standard Process for Data Mining (CRISP-DM) consists of six phases intended as a cyclical presented in Figure 1. In CRISP-DM, we assumed that knowledge discovered from data is a result of a process. The life cycle of this process consists of six steps:

- Business understanding;
- Data understanding;
- Data preparation;
- Modeling;
- Evaluation;
- Implementation.

The assumption of the model is that the analysis process cannot be carried out without understanding the context of business processes. (Kulikowski, (2015)) applied CRISP–DM as a model of conducting employees attitudes and opinion research. The author emphasized that CRISP– DM model through structuring and organization of the research process can improve research management and enable more efficient knowledge discovery from collected data. CRISP-DM model is seen as a comprehensive collection of tips and techniques in the analysis process, which results in the systematic and structured analysis (Kuligowski, (2015), p. 114). (Xiao et al., (2015)) recommend teaching of mathematics in the field of data modeling. (Venter, (2007)) presents a novel approach involving the adaptation of CRISP-DM, a cross-industry standard process for data mining, to CRISP-EM, evidence-mining methodology designed specifically for digital forensics. A similar approach is recommended by (Riedel et al., (2014)), (Rivo et al., (2012)) and (Khaleel, (2013)).

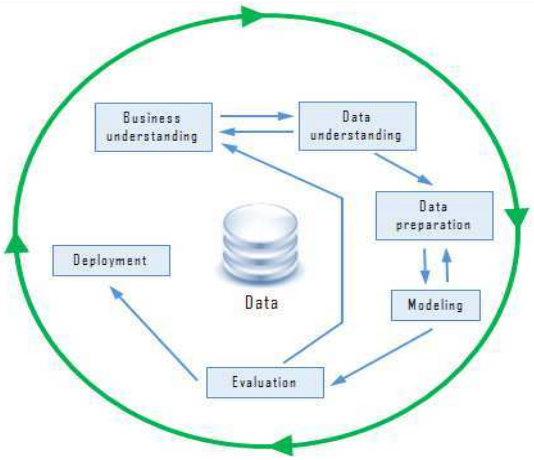

Fig. 1. Cross-Industry Standard Process for Data Mining
Source: www.zentut.com/data-mining/data-mining-processes [16.02.2017].

Our objective is to use the CRISP-DM model to extract knowledge from Twitter. However, we concentrated on the first two stages through the prism of the cultural context. The analysis process of data from social media is unusual. In Table 1, we provided details of CRISP-DM model tasks and then we described business and data understanding stages.

Choosing the right and clear path of modeling such data can have many advantages, such as:

- Systematization – which is needed for working unstructured data;
- A clear framework for research – with big data streams, researcher can fall into the (so-called) *trap of thinking*, because of the large number of potential information analyst will have a problem with the selection of appropriate variables. CRISP-DM model allows staying within the established framework;
- Establishing links between the organization's goals and the objectives of the study;
- Assessment of the possibility for the organization to implement such type of the analysis, such as budget resources, technical resources and personnel resources.

Business understanding

Twitter exploration should be designed to be useful to achieve business goals. The relationship between the analysis and business goals should be clearly defined. This is an important step to define what the organization (or researcher) wants to achieve by monitoring social media. To apply appropriate analysis, techniques must be clarified whether the business goals will require analysis of: (1) a Twitter trends (hashtags), (2) twitter network - topology, (3) events spread over the network, (4) identification of influencers or (4) analysis of sentiment. By answering these questions, researchers choose the appropriate techniques of analysis. By knowing analysis goals, they must define the task of data analysis and methods of achieving the goals set by the organization. The

task at this stage is to translate business needs into a data mining problem and explain the business situation that drives the project.

Table 1. The stages of the CRISP-DM model with assigned tasks

No.	Stages of the CRISP-DM model	Assigned tasks
1.	Business understanding	• Specifying business goals. • Assessing the context of analysis in organization's business processes. • Determining the goals for data analysis. • Creating a plan for data analysis.
2.	Data understanding	• Gathering data samples. • Describing data. • Forming data mining sample data set. • Verifying data quality.
3.	Data preparation	• Getting data set. • Data cleaning process. • Creating new variables.
4.	Modeling	• Specifying techniques of data analysis. • Creating criteria for evaluation the results. • Model building. • Evaluating models.
5.	Evaluation	• Evaluating results. • Evaluating the entire analysis process. • Determining if the process can be implemented or deciding to re-analyze. • Creating report.
6.	Deployment	• Creating implementation process plan. • Creating tools for monitoring. • Final report.

Source: (Kuligowski, (2015).

Data understanding

Data from Twitter is a specific data type. There are different types of Twitter data, such as user profile data and tweet messages. The former is considered static, while the latter is dynamic. Tweets may content text, images, videos, URL, or spam tweets. To track and monitor different events, most studies began with collecting the desired datasets from Twitter and applying filtering techniques to remove redundant data or spam tweets.

Thus the difficulties start at a very beginning: connecting to Twitter via Streaming API allows for creating datasets in a "real-time." Common (thanks to the experience of researcher) pre-filtration takes place at this stage, for example: the selection of a tweet language, tweets geolocation and definition of search criteria such as keywords or hashtags. Why is this an important step in the analysis? Because Rest API restricts access time for a messages to one week back, so obtaining information about certain past events is possible only at the specified time (Michalak (2016)), (Bonzanini, www.packtpub.com, (27.08.2016)).

Therefore, understanding data is very important in the case of Twitter because of the specific nature of data (limit of 140 characters, etc.) and the challenges posed – which are discussed in work: (Giachanou & Crestani, (2016)).

As we can see, cultural context is part of a *business understanding* and *data understanding* stage, which may or may not need to be considered, depending on the aims. Does the cultural context influence user feedback? Do you require additional criteria for filtering messages? Applications such as sentiment analysis? Are some messages more emotionally charged than others?

CULTURAL CONTEXT IN SOCIAL MEDIA

Previously, it was a challenge for marketing specialists, but also now not only does the information reach millions around the globe but it is also easier to reach the target audience in specific parts of the world. Having the ability to reach people worldwide has many advantages (Egros, compukol.com/social-media-usage-across-cultures, accessed on: 31.03.2017). This means that communication occurs across the borders or different countries, continents or even cultures. It is a desirable effect, however the social media user across the world differs in respect to their culture and data shows that significant differences exist in a way people in the world use social media.

High-context cultures and low-context cultures – E. T. Hall

Cultural diversity has a huge influence on the communication effectiveness in general (Mikuła, (2015)). Due to its range and speed, the message has a strong meaning for communication among the users of social media platforms. We can distinguish main differences in ways of communication according to high-context cultures and low-context cultures. E.T. Hall who presented this classification claims that the key to effective intercultural communication is to understand cultural differences that divide nations of the world (Gulak-Lipka, (2016)).

Table 2. High-context and low-context cultures according to E. T. Hall theory

Low context culture	High-context culture
Australia, South Africa, Austria, Belgium, Denmark, Netherlands, Finland, France, Ireland, Israel, Canada, Germany, New Zealand, Switzerland, Sweden, USA, UK, Italy	Saudi Arabia, Argentina, Brazil, Chile, Ecuador, Greece, Guatemala, Columbia, Costa Rica, Poland, Salvador, East and West Africa, Asian and Fareast Countries
Individualism - Promote resourcefulness and responsibility for your deeds	Collectivism - identification with the group, the large role of customs as carriers of culture
The role of verbal communication	The role of intuition in communicating
Monochronic culture	Polychronic culture
Time is of high value, appreciated punctuality, monochrome time - one thing is done at once	No respect for time, polychrome time - many activities are performed simultaneously

Source: own analysis based on (Hall, (2001)) and (Kamińska - Radomska, (2012)).

According to these criteria, we can describe how people communicate within these countries. Typically, in low-context cultures, the main part of the message is passed verbally using many words. As a result, communication is direct, clear, packed with many facts, details, specifications that make the message very accurate and understandable to the recipient. In contrast, in high-context culture societies, where it is not customary to speak 'straight from the shoulder', a large part of the information is transmitted between words using context. Sometimes the statements are quite vague with a lot of ambiguity and lack of direct information. This is due to the fact that other people assume that the range of information is obvious so there is no point in having to speak about it at all because it can be picked out from the context. In contrast, much more important in all communication is the tone of voice, gestures and facial expressions, which can change the meaning of the entire statement, and only on the recipient does it depend whether it will be decrypted correctly. However, this only applies to face-to-face communication. When using social media, we can make up for the situation by using the whole range of emoticons, gifs, exclamation marks or choosing verbal description of our mood, etc.

Cultural dimensions – G. Hofstede

Another way to look at cultural differences in social media communication can be through the cultural dimensions presented by Geert Hofstede. His national culture model consists of six dimensions. The cultural dimensions represent independent preferences for one state of affairs over another that distinguish countries from each other. Countries can be easily compared to each other and thus present a meaningful picture of cultural differences. The model consists of the following dimensions (geert-hofstede.com/ accessed on 13.02.2017):

- Power of Distance, where the fundamental issue is how a society handles inequalities among people. People in societies exhibiting a large degree of Power Distance accept a hierarchical order in which everybody has a place and which needs no further justification. In societies with low Power Distance, people strive to equalize the distribution of power and demand justification for inequalities of power.
- Individualism versus Collectivism: a society's position in this dimension is reflected in whether people's self-image is defined in terms of "I" or "we".
- Masculinity versus Femininity: the Masculinity side of this dimension represents a preference in society for achievement, heroism, assertiveness and material rewards for success. Society at large is more competitive. Its opposite, femininity, stands for a preference for cooperation, modesty, caring for the weak and quality of life.
- Uncertainty Avoidance Index: it expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity and how society deals with the fact that the future can never be known: should we try to control the future or just let it happen.
- Long Term Orientation versus Short Term Orientation: in the business context this dimension is related to as short term – normative versus long term – pragmatic approach.

- Indulgence versus Restraint: indulgence stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun. Restraint stands for a society that suppresses gratification of needs and regulates it by means of strict social norms.

Literature review – overview of conclusions

According to the above, in order to present the relationship between the culture and social media we need to look a little closer into a particular country or culture. In some cultures, countries social media and culture go hand in hand. For example, in the USA, Belgium and the Netherlands. USA, like with many other novelties, was the first country to adopt the social media. The Netherlands followed 2-3 years after the US, followed by Belgium, 3-5 years after the Netherlands. Adoption of new "*things*" in general, which also applies for social media, largely hinges on Uncertainty Avoidance. Low scoring countries adapt faster to new technologies than higher scoring cultures. The US has a relatively low score in this dimension, followed by the Netherlands, with Belgium having a relatively high score. The cultural dimension that plays a role in the Netherlands topping the list of Twitter users is Femininity (or a low score on Masculinity). The Dutch are consensus-seekers and in order to reach consensus, you need to talk and share what is on your mind. Hence the love for Twitter in the Netherlands (Smit, (2015)).

Interesting remarks can be made about the fact that the Belgians conduct their business in a different way than the Dutch and Americans (who are more similar). In Belgium, business culture relies more on relationships and it sort of proves why the penetration of e.g. LinkedIn and Twitter has not been as fast as in the US and the Netherlands. Generally, Belgians don't see the reason to need LinkedIn to connect with the people they already know. In fact, they prefer to do business with people whom they know (or are referred to) as opposed to doing business with total strangers. Doing business with total strangers takes by far longer to move forward in Belgium than in the other two countries. Some statistics support that. In the US and the Netherlands, over 75% of the HR professionals use LinkedIn to find new candidates for a job, while in Belgium the percentage is around 40 (Smit, (2014)). Smit also shared the result of a small research, showing that the majority of people in Belgium were not allowed to use social media at work for private purpose, however they still did, using their own smartphones. This can be attributed to the relatively high score on Power Distance, compared to the Netherlands and the US.

A great number of statistical data on social media use can be found in Global Social Media Research Summary 2016. For example, it indicates that East Asia is a great market for social media platforms and possibly for doing business through them since there is the biggest number of active social media accounts- 33%, which is the highest in the world. This could be helpful for anyone who is planning to do business in that part of the world. However, many parts of the world e.g. Pacific and Oceania also note a constant growth at a very impressive pace. When looking at the world by the number of Internet users, it really brings home the importance of the East Asian and South Asian markets. Digital communication offers new opportunities to reach these people, although as always, cultural differences are considerable challenges to international marketers.

Since there is such a big market for social media in East Asia, aside from the well-known and most popular social media platforms like Twitter and Facebook, people there have an option to choose their own specifically designed and structured social network, best suited to their preferences and culture. Some networks are closed ones, for example Mixi in Japan, whereas Twitter is a worldwide, global or just simply open network. In some Asian countries the most popular social media platforms allow their users only to play games with other users.

Anyone who is analyzing the content in social media, especially for scientific or business purposes, needs to be sure to collect correct data in order to achieve correct and not misleading results. One of the goals of this article was to explore if cultural context in any way reflects or has any influence on the way social media users communicate via them. The authors also attempted to find an answer to the question if there is a need to add any extra filtering on collected data in order to understand and correctly interpret the cultural context in social media users' posts. If we were to solely rely on Edwards Hall's theory of high and low context cultures, it would be fairly easy to perform most of the analysis. When it comes to western and eastern cultures, they appear to have totally opposite tendencies and lie on the opposite ends of the context scale.

On the very basic level, we can classify how and for what purposes nations/cultures from different parts of the world use social media. Semiocast has conducted a 'Geolocation analysis of Twitter accounts' and found for example, that in the Asia and Pacific region people either use social media to stay in touch with their friends (Australia and China), search for current news or events (South Korea) or even search for employment (India). Japanese users prefer to use social media as a real-time communication tool (Twitter). Additionally the most active users of Twitter in Japan (14.9%) are among the younger generation (age 15-19) and younger users view Twitter as a real-time communication tool. On the other hand, the older generation tends to use Twitter mostly to gather information. When compared to adolescents, they tweet less frequently. In Europe, users mainly search for posts with news and information on interesting events (Netherlands, Italy, Spain and Russia), keep in touch with friends (France) or even search for new products to purchase (UK). Americas also rely on social media when they search for products to buy (USA) and search for information, usually short instructive videos on how to do different things (Brazil).

So far, there have been very few studies analyzing the relation between social media and culture, but those few prove that the cultural context in this case is not so obvious. For example, Acar & Deguchi suggests that there is a subtle and complicated relationship between culture and Twitter use (Acar, Deguchi, (2013)). They made an attempt to fill the literature gap and promote cross-cultural understanding. For that purpose they conducted a study on 200 students from Japan and the US and their 4,000 tweets and formed the three following hypotheses:

- H1: Japanese users post fewer self-related tweets than Americans.
- H2: American users post more self-promotion related messages than Japanese.
- H3: American users ask more questions on Twitter than Japanese.

Table 3 shows key differences in the way of communication according to E.T. Hall's theory.

Table 3. Key differences in communication according to E.T. Hall's theory

Japan	USA
High-context culture	Low-context culture;
Communication ambiguous, not clear;	Communication direct, open;
Focus on keeping the harmony (no chaos) and group solidarity;	Based on honest and sincere intentions and feelings;
Unwilling to exchange information with others, only posts about others;	Concentrated on auto promotion and independence;
Asking questions is associated with requests and is not perceived well/welcomed in the Japanese culture.	Communication invites/requires asking follow-up questions (for better understanding, additional information)

Source: own work based on (Acar, Deguchi, (2013)), (Hall, (2001)).

The results of this analysis showed that Japanese college students post more self-related messages and in general ask fewer questions compared to American college students. This is a quite surprising conclusion especially since the US is rather an individualistic, self-oriented and self-reliant society. What is more, it has been found that tweets that refer to TV are more common in Japan, whereas sports and news tweets stand out in the US. This pretty much confirms what Sakaki, Okazaki, and Matsuo claim. According to them, all microblogging services have one thing in common: real-time information update. The simplicity and ease of use of microblogs like Twitter allow their users to post information even several times a day or more. Out of the three hypothesis, the authors rejected the first two that seemed obvious from the point of view of high and low context cultures.

CONCLUSIONS

There is a huge potential in getting more business opportunities by being culturally sensitive and knowing how to design products and services for different countries. Understanding why social media in different cultures is used differently is probably the most important for people working internationally, using social media surrounded by different cultures. Nevertheless, the link between Twitter social platforms and the misperception of the cultural perception is not clearly defined and discernible.
In summary, *we need to keep in mind* that:

- Cultural context may affect the way users respond - in terms of the way emoticons are used -> Emotional saturation of messages. -> Application: special attention in sentiment analysis.
- Anonymity in the network affects the way that users perform they opinions.
- Twitter is intended to express the opinion of the individual in relation to events - the opinion is dependent on a number of factors and can change quickly.

- Social media as a platform for communication without borders blurred ***may not*** be affected by cultural differences.
- Different social media platforms are popular in different countries - keep in mind that the online discussion may be more intense on another platform than the one that is the source of data.

The main recommendation is as follows:

- **Remember** cultural differences because they can explain certain relationships, such as the engagement and intensity of commenting of certain events. Remember about the cultural context of marketing campaigns, product design, and incorporate this aspect into the design phase of your research. However, due to the diverse nature of the description of the *social media in case of cultural context* in literature, it should be borne in mind that it requires further work and research.

REFERENCES

Acar A., Deguchi A., (2013), "Culture and social media usage: Analysis of Japanese twitter users", *International Journal of Electronic Commerce Studies*, Vol.4, No.1, pp.21-32, 2013.

Bokszański Z., (2007), Indywidualizm a zmiana społeczna, *Wydawnictwo Naukowe PWN*.

Bonzanini, M. Mastering Social Media Mining with Python, retrieved from: https://www.packtpub.com/ [27.08.2016].

Ebner M. & Schiefner M., (2008), Microblogging—more than fun? In: I. Arnedillo Sánchez & P. Isaías (Eds.), *Proceedings of the IADIS Mobile Learning Conference* (p155-159). Lisbon, Portugal: IADIA.

Egros, A., Social Media Usage Across Cultures, retrieved from: compukol.com/social-media-usage-across-cultures/

Giachanou, A. & Crestani, F., (2016) Like it or not: A survey of Twitter sentiment analysis methods*, ACM Comput. Surv.* 49, 2, Article 28.

Gulak-Lipka P., (2016), Intercultural management on the basis of a sports club, *Acta UNC Zarz*., Z. 43 No. 3, pp. 63-80.

Hall E.T., (2001), Poza kulturą, Warszawa, *Wydawnictwo PWN*.

Kamińska-Radomska I., (2012), Kultura biznesu, normy i formy, W*ydawnictwo Naukowe PWN*, Warszawa.

Khaleel M.A., (2013), A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases*, International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 12, pp. 149-153.

Kowalczyk S., (2010), Elementy filozofii i teologii sportu, *Wydawnictwo Katolicki Uniwersytet Lubelski.*

Kulikowski K., (2015), Zastosowanie modelu CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM) W badaniach postaw I opinii pracowników., *Zeszyty Naukowe Politechniki śląskiej, Seria: Organizacja i Zarządzanie z 82*, No. Kol. 1940, pp. 111-121.

Larose D.T., (2006), Data mining methods and models, Wileyo-Intesscience A John Wiley & Sons, *INC Publications*, retrieved from: https://www.researchgate.net/file.PostFileLoader.html?id=589233072l7e20ed3b2cc056 HYPERLINK "https://www.researchgate.net/file.PostFileLoader.html?id=589233072l7e20ed3b2cc056&assetKey=AS:457031996973056@148597632733"& HYPERLINK "https://www.researchgate.net/file.PostFileLoader.html?id=589233072l7e20ed3b2cc056&a

sset-Key=AS:457031996973056@148597632733"assetKey=AS%3A457031996973056%40148597632733

Michalak J., (2016), Detecting sentiment in Twitter data – supervised machine learning approach for Twitter Sentiment Analysis in Python, *Torun Business Review*, v 15, No. 4, pp. 97-110.

Mikuła B., (2015), Współczesne tendencje w zachowaniach organizacyjnych, *Uniwersytet Ekonomiczny w Krakowie*.

Riedel M., Memon A.S. & Memon M.S., (2014) High productivity data processing analytics methods with applications, "*Information and Communication Technology, Electronics and Microelectronics (MIPRO*), 37th International Convention, pp. 289-294.

Rivo E., De La Fuente J., Rivo Á., García-Fontán E., Cañizares M.Á. & Gil P. (2012) Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical and Translational Oncology,* 14, pp. 73-79.

Smit C., (2014), How to overcome cultural differences in business Avoid the Mistakes that Everyone Else is Making When Doing Business Internationally, *Amazon Digital Services LLC.*

Smit C., (2015), Uncertainty Avoidance in international business, The Hidden Cultural Dimension You Need to Understand When Doing Business Overseas*, Amazon Digital Services LLC.*

Venter J., de Waal A. & Willers C., (2007), Specializing CRISP-DM for evidence mining, [in:] P. Craiger, S. & Shenoi, S. (eds.) Advances in Digital Forensic III pp. 303-315, *Springer, Boston*.

Xiao X., Xu H. & Xu S. (2015), Using IBM SPSS modeler to improve undergraduate mathematical modelling competence. Computer Applications in Engineering Education, in Press.

about.twitter.com/company
dev.twitter.com/overview/apidobj
geert-hofstede.com
http://www.gartner.com/technology/home.jsp
twitter.com
zentut.com/data-mining/data-mining-processes
http://semiocast.com/en/