Text Mining & Social Media Data Opportunities & Challenges

By Ida Johanne Austad

ID: 10358736

COMP61332 Text Mining

By Ida Johanne Austad (10358736)

Introduction

Text Mining techniques can be applied in many domains and on textual data from many different sources. A rather recent source of enormous amounts of textual data is social networks. As social networks such as Facebook, Twitter, Linkedin and many others are becoming an ever present part of our lives and communication, both in our private lives, but also between us, public institution and businesses, this textual data source is getting an increased amount of attention. Different industries and actors in society are realizing that new knowledge and unknown patterns can be extracted from this data. With this great potential however, comes distinct challenges which differ from that of many other sources of textual data. This essay will discuss the opportunities of using text mining techniques for three selected actors in society, and then discussing the challenges which must be handled by them all to be able to take advantage of the fast-evolving text mining techniques without harming users' trust and ensuring reliable results. The aim of the discussion is to provide an overview of the variety in opportunities across actors in society, and the challenges than face – rather than go into detail about various techniques and technology.

What is text mining?

Abdous & He (2011) define text mining as techniques which aim to find models or patterns in unstructured data in the form of text. When such techniques are automated they can help us extract and exploit potentially new knowledge from texts in an efficient and systematic manner (He et al., 2013). It differs from information retrieval in that does not just concern extracting and organizing the text data, but it takes it several steps further by allowing for potential discovery of new knowledge.

Social media and textual data

Since the early 2000s the invention and use of different social media networks has increased significantly (Edosomwan et al., 2011). Social networks help people communicate and discuss topics with others, without being prevented by geographical distance. The different social medias have thus become an important platform where people learn from each other and share knowledge (Irfan et al., 2015, Sorensen, 2009). Due to these great opportunities the use of social media networks has become omnipresent in our everyday lives, and as such an enormous amount of data is created by its users all around the world. Using text mining techniques on the part of this data which is in text can allow us to reveal completely new knowledge about what people are talking about, their opinions, feelings and events in their life (Irfan et al., 2015). Although large amounts of data can be a luxury, text in social media is also known to be noisy, often written with lack of attention on correct grammar and spelling - which can result in lexical, syntactic, and semantic ambiguity (Barbier & Liu, 2011, Salloum, 2017). Moreover, the text is accompanied by hashtags and emoticons which must often be included in the analysis to understand the meaning of the text. All in all, although social media textual data comes in large quantities, it can be more challenging for a text mining technique to decipher it's meaning, compared to many other written resources online.

Opportunities

The massive adoption of social media has resulted in the application of text mining techniques within several areas of society and for various purposes. The examples which are selected for discussion here are applications for businesses, emergency services and public health care, to emphasize the wide specter of use cases for text mining techniques in the present and future.

For businesses, especially in the B2C (Business to Consumer) market, analyzing the textual data generated on their own and their competitors social media profiles can reveal hidden knowledge which was either practically impossible to reveal before or very costly; through extensive market studies and customer research (He & Zha, 2013). Marketers can now use text mining techniques to get a better understanding of what their competitors are doing, and how their customers perceive themselves and their competitors. Through acquiring this new knowledge and patterns they can get closer to obtaining a competitive advantage (He & Zha, 2013). In addition to the more high level understanding of their market, businesses could also be able to communicate more efficiently with their customers. One example is how businesses increasingly use social media to perform surveys, pull feedback and opinions from their customers. Using text mining tools on this data can make it much easier for the businesses to reveal patterns in their customers' enquiries, and as such find root causes, relations between a large number of enquiries or help them solve the right issues first to make their customers more satisfied (He & Zha, 2013).

The opportunities for using social media textual data to enhance emergency services ability to understand a situation and make more optimal decisions in times of crisis and mass emergencies has also become increasingly investigated. For instance, by using Twitter textual data one can immediately get first-hand information from the people who are on the scene where an emergency is taking place. Amongst others Yin et. al. (2015) and Zielinski et al. (2013) suggest that using text mining techniques to analyze the large amount of real-time textual data generated by social media platforms like Twitter in cases of mass emergencies can allow for emergency personnel to understand the impact of hazards and in turn act faster and in a more informed manner. In addition to investigating the ability of insights generated at the immediate time of event, resulting in quick, meaningful actions, Huang & Xiao (2015) argue that the analysis of social media textual data generated both before and after an emergency or crisis can help emergency personnel improve their ability to prepare for and organize their recovery initiatives before and after a crisis. Moreover, it can be used to revise and improve emergency response plans, which may be a much more cost effective way to obtain people's thoughts and experiences compared to traditional survey techniques (via phone calls etc.).

Public health care is another area where the applications of text mining techniques on social media data has been given attention, hoping to unravel insight which is difficult, time-consuming or sometimes impossible through traditional data sources such as clinical encounters. Amongst the use cases are for instance the ability to monitor (also known in biosurveillance) the outbreaks of different diseases and health threats such as influenza, food poisoning and chemical contamination. An example of how such insights could be used is to improve responses to disease outbursts, for instance to decide where to send vaccine supplies first - ensuring that the areas which are the most affected is covered as soon as possible (Dredze, M., 2012). In addition to surveillance, Dredze, M. (2012) argues that analyzing social media data can be used to support health risk assessment. By analyzing what people post and comment regarding topics such as smoking, alcohol, consumption, exercise etc. one can create different hypothesis which would potentially not been thought of otherwise due to unknown connections in the data. Especially since patients may be reluctant to share certain details and thoughts on their condition and how they handle it, using social media data can give health services access to knowledge about people's behavior, opinions and self-medication which they would not get through traditional medical appointments or surveys. Also, subpopulations which it is not as easy to get in contact with by traditional means could be reached when analyzing textual data from social media.

Challenges

Although we have seen from the discussion above that the opportunities and use cases vary across industries and actors, the challenges they must handle largely boil down to be the same. Still, the severity and "show-stopper" potential of the different challenges can vary. A selection of these will be discussed in this section.

First and foremost, as mentioned previously, the text which people post in social media is different from a lot of other textual data on which text mining techniques is often applied. Firstly, text from social media is often noisy, and does not necessarily adhere to rules of spelling and grammar - leading to difficulties when analyzing these texts due to the increased potential of ambiguity (Sorensen, 2009, Zielinski, A. et al., 2013). Secondly, the language used in social media is dynamic, and which words are used for certain events, emotions etc. change over time (Irfan et al., 2015, Yin, J. et. al., 2015). New words arise amongst people in their daily lives, which will often not be made use of in more formal written medias and documents, but are more likely to be used in social media. As such, it may be hard to train models to capture the essence of new texts when the essence of a post or message is reflected in an expression which is new, both to the world - but even more so to text mining technique which will often require to have observed these words before to be able to make sense of its meaning and which words it is associated with. Overall, these characteristics on social media data makes it more complex to analyze and thus to discover unknown patterns.

Another challenge, which is especially relevant to applications where success rely on situational information, such as the use cases discussed for emergency services and health care when it comes to responding better to disasters and disease outbreaks, is people's tendency to not include geotagging in their posts. Users may often avoid using geotagging, for instance to protect privacy (Yin et. al., 2015). If this is the case then one may have to rely on which locations are mentioned in the post and in hashtags. However, it is not uncommon that people far away from an occasion to share posts which are related or not to the event of interest with such mentions or tags. The result is that one either has to create a model or technique which is able to filter out location mentions and tags which are not relevant or reliable for the use case, or just rely on posts which have a geotag. The latter may result in not enough data to train the model for a specific application, or in general to decide which action is appropriate.

A third challenge which is common across many use cases for text mining on social media is the potential of bias. It has been found that certain groups in society, such as households with low income, low education or the elderly, may not have the motivation or skills to learn how to use social media (Huang & Xiao, 2015). Thus, one has to consider the consequences of the fact that social media users does not represent the entire population (Dredze, M., 2012). For instance, Huang & Xiao (2015) suggest that in case of emergencies certain areas may not be able to use social media because their area being too severely damaged, which again could cause a model based on social media text to not detect the need for aid in this area. Moreover, in the case of health surveillance, different groups are expected to share information about a disease or lifestyle habit on social media, while other groups may be more likely to report an inaccurate diagnosis (Dredze, 2012).

The last challenge to be highlighted here is the issue of privacy. That those who implement and make use of text mining techniques base their action on ethical values is crucial according to Irfan et al. (2015). This is the only way that the community will trust these actors and their solutions. Although the data posted on social media is public, users expect a certain level of privacy (Dredze, 2012). This has to be taken into consideration as one can even infer unstated facts about the people based on what they post, such as demographics or diagnosis (Dredze, 2012). With the rising focus on privacy, at least in certain parts of the world - such as in Europe with GDPR, the need to take special care when handling personal information, although publicly shared, increases in importance to maintain trust.

Conclusion and thoughts on the future

As the discussion in this article has shown the applications across actors and industries for text mining techniques are different. Some propose value to the public and some propose value to businesses and institutions, or both. Still, the challenges which they face are often the same, and there are many more which have not been discussed in this article. What seems to be certain is that if text mining techniques can be applied in a manner which overcomes these challenges it has the potential of improving different parts of our life and society in a substantial way.

References

Abdous, M., He, W., & Yen, C. (2012). Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade. Educational Technology & Society, 15, 77-88.

Barbier, G., & Liu, H. (2011). Data mining in social media. Social Network Data Analytics, 2011, 327–352.

Dredze, M. (2012). How social media will change public health. IEEE Intelligent Systems, 27(4), 81-84.

Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. (2011). The history of social media and its impact on business. Journal of Applied Management and entrepreneurship, 16(3), 79-91.

He, W., & Zha, S. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. Int J. Information Management, 33, 464-472.

Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. ISPRS International Journal of Geo-Information, 4(3), 1549-1568.

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., Kolodziej, J., Wang, L., Chen, D., Rayes, A., Tziritas, N., Xu, C.-Z., Zomaya, A. Y., Alzahrani, A. S. and Li, H. (2015) “A survey on text mining in social networks,” The Knowledge Engineering Review. Cambridge University Press, 30(2), pp. 157–170. doi: 10.1017/S0269888914000277.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J, 2(1), 127-133.

Sorensen, L. (2009). User managed trust in social networking - Comparing Facebook, MySpace and Linkedin. 2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 427-431.

Yin, J., Karimi, S., Lampert, A., Cameron, M.A., Robinson, B., & Power, R. (2015). Using Social Media to Enhance Emergency Situation Awareness: Extended Abstract. IJCAI.

Zielinski, A., Middleton, S.E., Tokarchuk, L.N., & Wang, X. (2013). Social media text mining and network analysis for decision support in natural crisis management. ISCRAM.