Data Quality Challenges in Twitter Content Analysis for Informing Policy Making in Health Care

Axel J. Soto
Dalhousie University
Halifax, Canada
soto@cs.dal.ca

Cynthia Ryan
EY
Halifax, Canada
cynthia.ryan@ca.ey.com

Fernando Peña Silva
EY
Halifax, Canada
fernando.penasilva@ca.ey.com

Tapajyoti Das
LeadSift
Halifax, Canada
tdas@leadsift.com

Jacek Wołkowicz
LeadSift
Halifax, Canada
jacek@leadsift.com

Evangelos E. Milios
Dalhousie University
Halifax, Canada
eem@cs.dal.ca

Stephen Brooks
Dalhousie University
Halifax, Canada
sbrooks@cs.dal.ca

Abstract

Social media platforms and microblogs have become popular fora where the general public expresses opinions and concerns on a variety of matters. As a result, private and public organizations have been looking into ways for finding, understanding and communicating insights extracted from this massive amount of text-based interconnected data. There are, however, important difficulties associated with the noisiness and reliability of the content that hinder the analysis of the data. This paper reports the main challenges found in a real-world experience with social media used as a source of data to support policy making and assessment. We also propose a set of strategies for the precise retrieval of data, the profiling of social media users, and the involvement of policy makers in the analytical process.

1. Introduction

During the last decade microblogs, and Twitter in particular, have become popular platforms for users to broadcast ideas, feelings, facts and opinions. Due to their massive use and open policies for access to the data, they can act as social barometers that measure public concerns and satisfaction on a wide range of issues. As a result, governments and private organizations have been exploring ways of tapping into the data to extract insights for policy making and assessment [1-3].

Social media is an exemplary representative of the inherent challenges of Big Data [4, 5], where vast amounts of data are published at an ever-increasing rate. Several research efforts have focused on the design of computational methods to digest and analyze large amounts of data, including social media data [6, 7]. However, despite the abundance of data mining methods available in the literature, there are several challenges when it comes to the analysis of noisy and semi-structured data [8], i.e. textual (unstructured) combined with typed or vocabulary-controlled data (structured). Given the universality of access to social media, there are also varying levels of value and interest in the data, so identifying the right subset of data for analysis is the first stumbling block, yet commonly overlooked [5].

Data quality is a critical area in data analytics and in its absence it would produce meaningless results regardless of the analytical methodology employed. This is particularly important in social media given the noisy conversational nature of the text content and inconsistent user reliability [9, 10]. In addition, most approaches focus on the data analysis aspects in isolation of other related tasks—from data acquisition and cleaning to result presentation via analysis and exploration of data, which is known as the entire *data science process* [11]. As a result, most organizations also seem to lack the capacity to effectively perform the entire pipeline of analytical actions [3].

This article is grounded in a real case experience that brought together researchers from industry and academia to harness the benefits and challenges of social media data for informed policy decision making. Our goals were to analyze the opinion of people towards different topics in the health care system (e.g. emergencies, home care, outreach for specific diseases, etc.). Yet one typically overlooked aspect in social media analysis that we address in this paper is the quality and accuracy of the data collected for analysis.

URI: http://hdl.handle.net/10125/49983
ISBN: 978-0-9981331-1-9
(CC BY-NC-ND 4.0)

We illustrate this aspect by restricting the data to the opinion of people living in a specific geographical area. A second contribution of our approach is to show how policy makers can be active participants of the analytical process, as opposed to they being mere consumers of an automatic black-box process.

The description of the approaches applied in this paper are organized around the main data and information quality challenges that we identified. While our experience was focused on the health care domain, the process hereby described can be generalized to other domains and practitioners interested in social media data analytics. The next section describes the methods applied including a revision of previous related work in the literature. Results and outcomes of our use cases are described in Section 3, while concluding remarks and open areas of research are discussed in Section 4.

2. Methodology

Data science is an ill-defined concept referred to as a family of inter-disciplinary approaches aimed at extracting insights from data [11]. The traditional paradigm considered this as a linear sequence of stages, where data is fed, cleaned/preprocessed, analyzed and output in a pipelined-fashion. However, modern approaches considered this process as an iterative one, where stages tend to require some iterative refinement [12]. Therefore, the approach taken in this experience in the context of our social media analysis for supporting policy making can be graphically summarized as depicted in Figure 1. Each stage in the diagram is accompanied by a question that summarizes its main challenge. This article focuses on the data quality issues found and the approaches taken during the data processing-intensive stages, which are indicated within the dashed rectangle in Figure 1. Twitter was selected as the social media platform due to its high popularity and the availability of an API[1] for obtaining *tweets* in real time.

2.1. Data collection: finding the right content

Extracting valuable insights requires having the right data in place or the means for retrieving this data accurately and efficiently. The main challenge for an information retrieval method is to collect textual data with high precision (collected documents are relevant) and with high recall (relevant documents are collected). This challenge is exacerbated in Twitter for three reasons. First, the shortness of posts combined with data quality issues, which include informal expressions, misspellings and ad-hoc terms, make the matching and retrieval of text more difficult. Second, Twitter's streaming API, which is the access to the stream of tweets, has a rudimentary programming interface that requires the provision of a set of hand-picked keywords (or user names) that act as a disjunctive—OR-based—boolean query. Finally, there is also a limit that at any time the retrieved results are capped at 1% of the overall Twitter traffic volume. However, given the enormous number of tweets posted worldwide, this last restriction does not represent a limitation for most analytical cases.

Originating these keywords is not only cumbersome but it is also difficult to determine their appropriateness. Selecting too few keywords may potentially leave relevant content out of consideration and has the risk of biasing the analysis to the particular subset retrieved. Selecting too many keywords may introduce spurious content that, if not identified, will distort or obfuscate the analysis. Previous efforts in the literature have looked into query expansion approaches using pseudo-relevance feedback and Twitter specific features for the accurate retrieval of Twitter data [13, 14]. Some recent works have proposed methods that apply filtering techniques in a user-supervised fashion [15, 16], which is the type of strategy applied herein.

2.1.1. Our approach. In general, it is preferable to favor recall over precision by collecting more tweets at the risk of also obtaining non-relevant content. In this way a postprocessing filtering can still be applied. On the contrary, tweets that are not collected at the time they are posted, are difficult to crawl later by means of the Twitter API, and hence will be likely not retrieved.

Let us assume that we have p different pre-defined topics of interest for our analysis. Each topic requires its own set of keywords, namely: $K = \{\kappa_1, ..., \kappa_p\}$, so that $\kappa_i = \{k_{i,1}, ..., k_{i,|\kappa i|}\}$ with $1 \leq i \leq p$ contains the keywords corresponding to the query of topic i. In our use case we devised these keywords in an iterative manner. At the beginning a few seed keywords were manually chosen for each topic of interest based on domain knowledge. A semi-automated process was then started where keywords for each specific topic were prompted by the system and the user could add them or not to each κ_i. This was done by computing a score for each keyword with respect to a topic i, i.e. $score(k,i) = f(k,i) \cdot itf(k,i) \cdot type(k)$, where $f(k,i)$ is the frequency of a keyword k in the retrieved tweets for topic i, $itf(k,i)$ is the inverse topic frequency, i.e. an inverse measure of how frequent the keyword is in the p topics, and $type(k)$ is a boosting factor that depends on whether the keyword appears as a hashtag, user mention or regular term. This strategy was particularly

[1] https://dev.twitter.com/streaming/overview

useful to identify relevant hashtags or institutional accounts that were not included from the beginning.

The query submitted to the API is the union of the keywords for each topic, i.e. $Q = \cup_{i=1...p} \kappa_i$, where the tweets corresponding to each topic are separated offline after the crawling. An important aspect is that the crawling process was monitored for its performance, which closes the loop indicated in Figure 1 and leads to the subsequent removal or insertion of new keywords. This monitoring process is further described in Section 2.3.1.

While calculating recall is not practically feasible due to the large number of tweets, precision can be estimated by random sampling and manual inspection. Given a sample of tweets retrieved for topic *i*, we computed precision as the ratio of relevant tweets to all the tweets in the sample. Depending on the specific topic, precision ranged from 80% to 97%. This of course depended on the amount of keyword refinement, so we typically stopped tuning the keywords when a precision above 80% was obtained. Yet, as topics evolve over time a weekly monitoring was necessary to keep precision at acceptable levels.

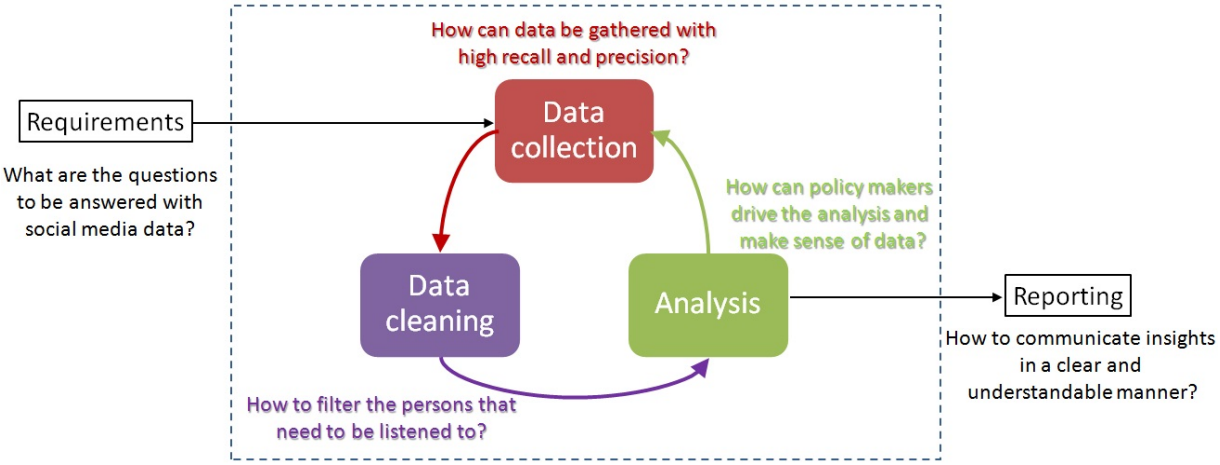

Figure 1: Data science stages and challenges identified in the context of social media analysis for policy making and assessment. Data-processing-intensive stages are within the dashed rectangle.

2.2. Data cleaning: finding the right users

Commonly, policies and decisions to be made are relevant to a specific group, such as people living within a specific geographical area, within a certain age or socio-economic group, or based on marital status. Identifying the opinions from specific groups of users is another major challenge, because the necessary information required to partition users is rarely made available as part of the public user profile in a structured form. For instance, less than 1% of the tweets contain geo-tagged data [17], while Hecht et al. [10] estimated that 34% of users provide fake or sarcastic location information in their own profile. Therefore, in order to have a large and reliable sample of data for analysis, the user profile information needs to be inferred or contrasted by other means.

There are two main types of methods for inferring user profiles. The first type looks into what can be inferred from the user's network of friends and the people that users most frequently or recently communicate with [17-19]. The other type looks into analyzing content or style of posts to infer user's profile data, such as age and gender, as it is done in [20], or by finding geographic references or regional language style to determine user's location [21, 22]. Other recent methods have looked into hybrid approaches that jointly exploit both network and text information [23, 24].

Another related aspect to user analysis is the fact that some users may have a higher impact when they express their opinions. Such users are commonly referred to as *influencers* [25] and they could be celebrities, press media members or just members of the general population with a large number of followers. Identifying influencers is key when analyzing social media content, as influencers' posts could be weighed higher since their messages are likely to reach more people, who are in turn more likely to trust in their content. Also, when content needs to be disseminated, influencers may be the target of directed messages with the expectation that they will engage in the conversation, and hence help broadcast messages to wider audiences. This practice facilitates improving "*corporate dialog*" [26]. Influencer

characteristics may vary depending on the thematic topic and scale, and as a result several studies have looked into their identification and analysis [27-29].

2.2.1. Our approach. In our use case we were interested in the opinion of people living in a particular province or state only. Given that the volume of geo-tagged tweets from that province about our topics of interest was minimal, inferring user profiles was critical to avoid having an insignificant number of tweets for analysis. The algorithmic inference we used is the result of an in-house supervised machine learning method [30], which combines different sets of features from different sources. It is worth noting that these methods were trained in a much larger dataset containing information from thousands of different users, and not just on the data crawled for this experience.

The first set of features was extracted from the text-based user profiles. This was done by running a Named Entity Recognizer [31] on the textual profile of each user. Extracted location entities were matched against a location database (http://www.geonames.org). Ambiguous cases can arise from people indicating more than one location, or from entities that could refer to multiple locations (e.g. "New York" could refer to the one in the state of New York or to the one in the state of Missouri), or by bogus locations (e.g. "Earth", which if not caught would be mapped to "Earth, Texas"). The second set of features were geo-tagged tweets, which are used to query the location database with the longitude and latitude data. While this adds precise information about the location, it is important to keep in mind the low percentage of users that enables the geo-tagging of tweets. In addition, avid travelers may have this information available, which forced us to only rely on geo-tagging if a large proportion of tweets are from a same location. Finally, the third set of features looked at information from the immediate neighborhood of the user. If the majority of mutual friends (followers and followees) are from the same location, then there is a high probability for the user to belong to that location too.

In addition, we analyzed how influential each of our identified users were. While there is no unique way to identify who or what makes a person an influencer, we developed a statistical approach that looks at a combination of features such as: number of followers, number of tweets, recency of tweets and promptness to engage in conversations. The algorithm performs as an outlier detector, in such a way that users who have feature values that are significantly higher than the mean feature values are considered influencers. It is worth highlighting that the more data crawled from Twitter users, the more reliable these statistics are.

2.3. Analysis: involving policy experts in the analytical process

In many data analytics scenarios the path from data to decisions is unclear and not everything can be automated. Typically, answering an analytical question leads to further questions about the data, so an exploratory or investigative analysis on the data is necessary. Domain experts, or policy makers in our case, should be involved in the analysis since they can pose insightful questions on the data and interpret situations better than data-mining experts who may not be familiar with the domain. Yet data experts need to be supported by data analytics/mining methods, which are able to crunch large amounts of data and provide descriptive or predictive statistics on the data.

Most of the challenge resides in designing appropriate and effective visual interfaces that can represent data in a useful form, and in developing machine learning and data mining algorithms that learn from data and from user intervention. This type of analysis is commonly referred to as visual analytics [12, 32], where interactive interfaces connect seamlessly with the data mining back-end. This gives more flexibility to the analyst than a monolithic solution that needs to be customized by the data mining expert every time data or analytical needs change. In the case of data analytics for policy making, it becomes important to support provenance [33]—the capacity to track interaction and changes on the data—and interpretability, so that the reasons behind the decisions can be understood and communicated. Therefore, methods that favor interpretability are preferred over black-box approaches that do not allow meaningful interaction. Tools for the visual analysis of semi-structured, interconnected and social media data have been presented before [16, 34, 35], but there are still vast opportunities for research that connects and expands upon these individual efforts from different areas.

2.3.1. Our approach. We developed two interactive visualizations that are aimed at exploring data in social media. The Javascript-based visualization library D3 was used for the development of our visual tools (https://d3js.org/) The first one is a simple interface that aims at discovering the main topics of conversation for a given time period as illustrated in Figure 2. Summaries of the conversations are shown as keyword clouds, where keywords are n-grams (*n = 1, 2, 3*) that are extracted based on the highest tf-idf values [36]. In this model, a *document* was considered to be the concatenation of all tweets happening in a

given time window. Note that if the most frequent terms were used as keywords instead of using the tf-idf values, it would show mostly the keywords used for crawling without exposing hardly any differences in each time window. At the right hand side there is a table of the tweets found for the given time period, where for each line it is shown the date and time, the name of the user (also known as *screen name* in the Twitter jargon) and the text itself. The tweets in the table can be sorted by any of these three features.

The interface allows the user to customize how many keywords are shown, the cardinality of n-grams considered for the keyword extraction, and the length of the time period. Upon interaction with the keyword cloud, the user can see a clicked word in a different color in the table of tweets. This allows investigation of the term in context and analysis of a narrower set of tweets. For instance, in Figure 2 we can see that the story of a girl rejected to hospital admission emerged in social media conversations on that day, which also led to other conversations of people demanding better policies regarding mental health issues.

This interface was used to monitor the quality of the retrieval and adjust the query when too much noise was perceived. This reinforces the idea, which is exposed in the cyclic structure of the diagram in Figure 1, where the analysis phase can lead to the improvement of data collection and cleaning. Yet in order to facilitate this cyclic process, domain experts should be involved in the analysis and have the right tools to explore the data.

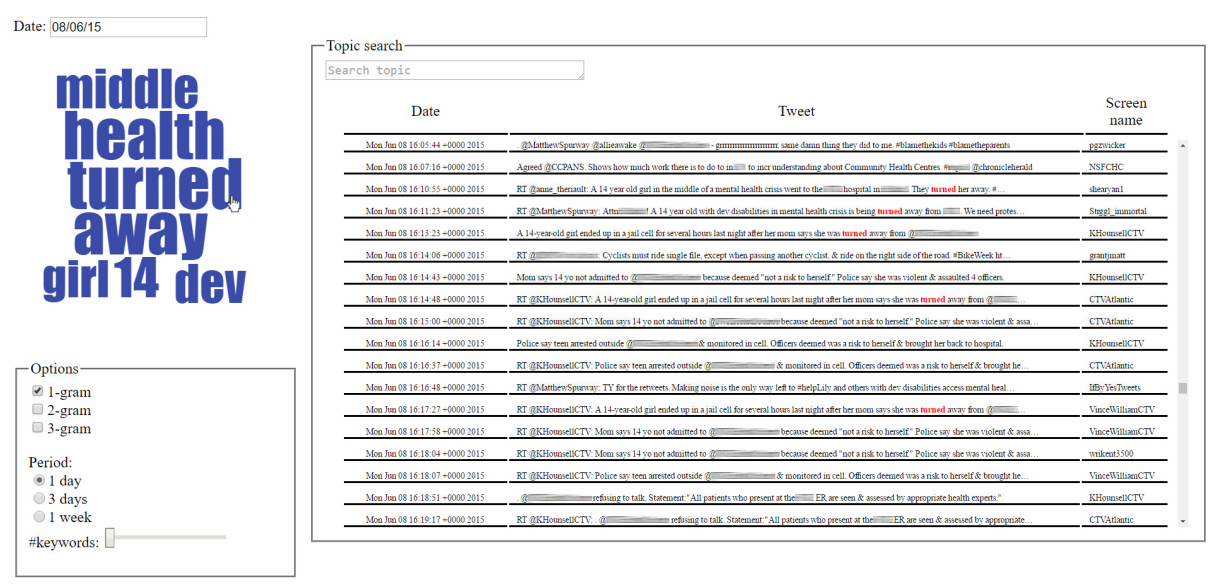

Figure 2: This interactive visualization allows the identification of the most meaningful stories for a given time period. If tweets are found to be off-topic, then action needs to be taken to refine the data collection and cleaning stages.

The previous interactive visualization, depicted in Figure 2, was appropriate to explore trending keywords in a given time span and provided a way of linearly enumerating the tweets containing those keywords. However, it had limitations when it comes to the analysis of interactions among users and their relative importance based on their network of followers.

Our second interactive visualization displays retrieved tweets using different visual encodings. For instance, Figure 3-a shows tweets from two topics of interest, namely "diabetes" and "home care", which are represented as circles in yellow and blue, respectively. Links between tweets indicate a reply or retweet. These circles and links typically exhibit a "star" network pattern, i.e. several tweets replying to a same tweet. The size of each circle can encode either the importance of the tweet, i.e. using the number or retweets or favourite counts, or user's importance, i.e. using the number of followers as its representative measure. Alternatively, tweets can be arranged temporally: from left to right encoding the date and from top to bottom encoding the time of the day, as it is visualized in Figure 3-b. When comparing the volume of posts about diabetes and home care, we can note that home care has become a more meaningful topic during 2015 with three main bursts of Twitter

posts (blue circles), while diabetes appears to be a topic that was more prominent during 2014 (yellow circles). In practice, the analyst should also consider any trend—in the case of Twitter is typically increasing due to the increasing number of users—to compensate for the difference in the absolute numbers of posts. As users can feel easily overwhelmed with the number of tweets, there are also filtering mechanisms to hide tweets based on user's importance, tweet's importance or recency.

3. Results

In this section we present the main results of the methods presented in this experience as well as the qualitative analysis of the exploratory tools. Twitter privacy policy and confidential agreements with the target healthcare organization do not allow the publication of the tweets used during the analysis. While this imposes some limitations in the sharing of research results, our experience also showed us that social media datasets should be

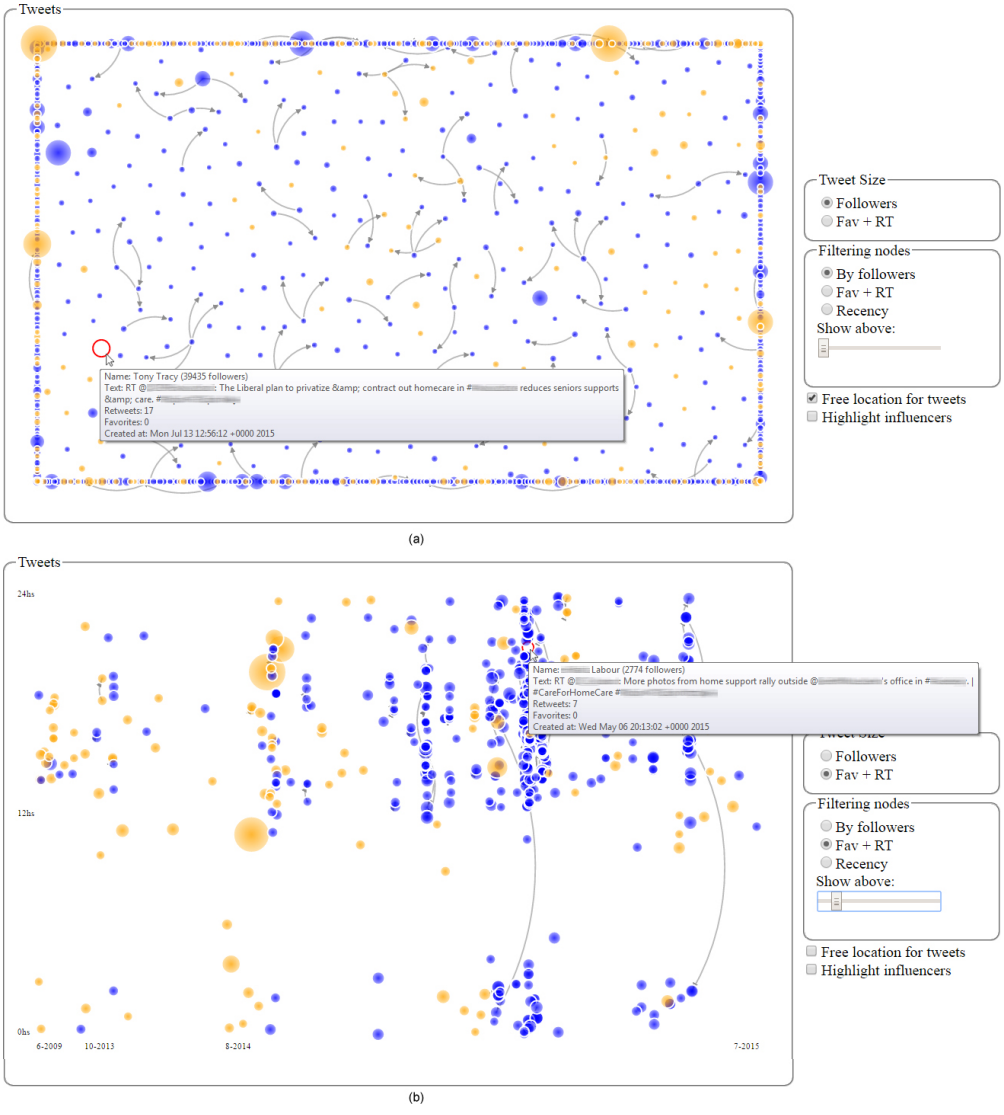

Figure 3: This visualization shows tweets as circles, where their size encodes their (user's or content) importance, and their color encodes the topic (home care and diabetes in this case). Links show replies or retweets among posts. (a) Tweets can be clicked, dragged, explored and filtered on demand based on tweet importance or posters' influence. (b) Tweets can be organized in chronological order to identify temporal thematic patterns.

evaluated in the specific settings and context under investigation, rather than expecting the results or labeling of other datasets being representative in other domains.

After contrasting our predicted locations with manually annotated locations via crowdsourcing (http://www.crowdflower.com/), we analyzed the accuracy of our approach at different levels of localization and in terms of the different features considered. As we can see from Figure 4, around 60% of the users could be geo-localized by using the profile information. However, less than 30% could be narrowed-down to the city level. The incorporation of GPS-tagged information allowed increasing the coverage of geo-localization by a 10%, and the city-level geo-localization increased around 15%. Incorporation of the inferred locations of the user's neighborhood allowed inferring the location of more than 90% of our sample, where 70% could be localized to the city-level.

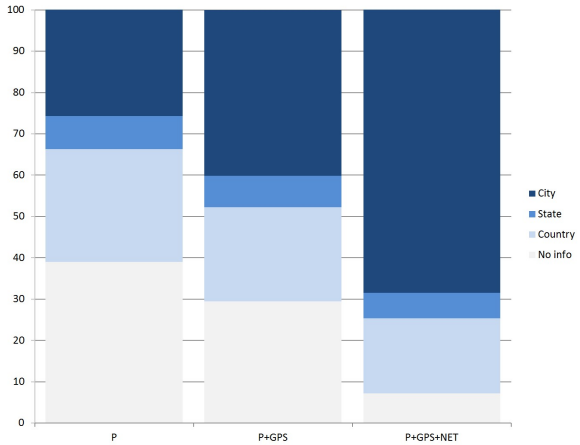

Figure 4: Evaluation of user multi-level (country, state/province and city) localization in terms of different types of features used. On the horizontal axis, P indicates that the localization method only uses the self-provided profile information, P+GPS when geo-tagged tweets is added, and P+GPS+NET denotes when the localization of mutual followers is considered.

The influencer classifier allowed an analysis of the percentage of influencers that our target health organization was either following or being followed by (Figure 5). This analysis revealed valuable information as almost 60% of the provincial influencers that were broadcasting opinions in health-related topics were not being read or interacting with this health organization. While the details of our predictive models have not been made available due to proprietary reasons, the research community can still benefit from these models by accessing LeadSift's API [30], which can be used for automatic user profiling of any Twitter user or as a benchmark for comparison with other approaches.

The tool for visual analysis of Twitter content received significant interest among health care policy makers, who lacked technical skills to collect and clean the data. Therefore, the main advantage they identified was the provision of a tool where they could focus on the data analysis, rather than on the technical challenges of retrieving information from Twitter and filtering the pieces that are relevant for their analysis. They also praised the fact that results are not static images but allowed some degree of interactivity that let them focus and highlight different aspects of the data. A tangible outcome of the experience hereby described was the analysis being forwarded to the groups in charge of diabetes to support their outreach programs. Such analysis helped triggered more analytical questions, such as the replication of the analysis on different states/provinces with similar demographics so that differences and commonalities in the data could be contrasted.

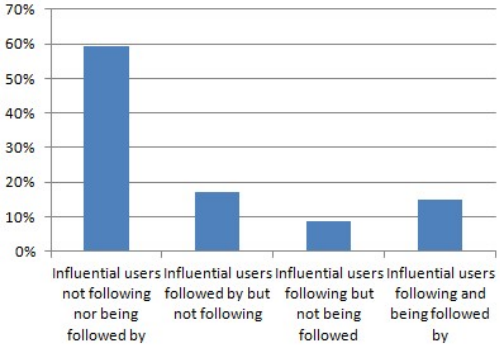

Figure 5: Analysis of followers and followees of our target health care. All the influencers are from the same specified geographical location.

4. Conclusions and open directions for research

This paper presented our experience in addressing data quality issues in social media data in the context of policy making for health care. Our main contributions of this article are: 1) the review and discussion of the main data quality challenges in social media analysis, 2) the consideration of simple, yet effective, approaches that consider the entire data analysis process—from data collection to analysis via cleaning—, 3) the proposal of exploratory analyses as a means to involve domain experts in the analytical process, and 4) the provision of an API that the research community can benefit from to enrich Twitter user profiles.

The main advantage of using Twitter data is the ability to obtain recent opinions and information in a cost-economical and unobtrusive way. Yet analyzing this data naïvely is likely to produce meaningless or biased results that are of little value to policy makers. While data bias appears to be an inherent data quality issue in the Web [37], we presented a set of strategies that addresses three major challenges in data quality for social media analysis. These strategies can be generalizable to other analytical tasks performed on social media data.

The first strategy addressed the problem of social media data retrieval by performing a continuous monitoring of the data crawled in order to keep it *precise*, and a term scoring approach to suggest keywords to be added to the query, and thus increase *recall*. The second strategy addressed the challenge of automatic user profiling by the application of machine learning models operating on different sources of data. This allows a better characterization of users, and hence a finer-grained analysis on the data collected. The final strategy aims at involving policy makers as part of the analytical process. This was performed by means of interactive visualizations that allow the experts to explore the data as a way of extracting insights and formulating new questions.

There are several open challenges in the area of data quality in social media as well as exciting research opportunities as a result of recent research findings in related areas. Although research on natural language processing and text mining can be regarded as mature, it has largely focused on the assumption of well-written "long enough" documents [36]. Several methods for word and phrase similarity have been proposed in recent years [38, 39], which aim at measuring similarity between text that may not contain any words in common. While not all these methods are directly applicable for social media, they represent promising approaches to identify related text content. However, other more advanced tasks that allow the understanding of and insight extraction from a large number of posts such as topic extraction, summarization or entity recognition have not been fully solved yet in the context of social media data [40, 41] and are worth further investigation. Finally, the mining of social media data by public and private organizations also opens a range of new challenges not necessarily related to its analysis. These challenges relate to privacy, accessibility and social inclusion and data governance in general [2], which also require their own study and strategies.

5. Acknowledgements

The authors would like to thank NSERC for support through an Engage grant.

6. References

[1] Chun, S.A., Shulman, S., Sandoval, R. and Hovy, E., 2010. Government 2.0: Making connections between citizens, data and government. Information Polity, 15(1), p.1.

[2] Bertot, J.C., Jaeger, P.T., Munson, S. and Glaisyer, T., 2010. Social media technology and government transparency. Computer, 43(11), pp.53-59.

[3] Kwon, O., Lee, N. and Shin, B., 2014. Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Management, 34(3), pp.387-394.

[4] Lazer, D., Kennedy, R., King, G. and Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. Science, 343(6176), pp.1203-1205.

[5] Tufekci, Z., 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In Proceedings of the Eigth International AAAI Conference on Weblogs and Social Media. The AAAI Press, 505–514.

[6] Aggarwal, C.C. and Zhai, C. eds., 2012. Mining text data. Springer Science & Business Media.

[7] Leskovec, J., Rajaraman, A. and Ullman, J.D., 2014. Mining of massive datasets. Cambridge University Press.

[8] Kandel, S., Paepcke, A., Hellerstein, J.M. and Heer, J., 2012. Enterprise data analysis and visualization: An interview study. IEEE Transactions on Visualization and Computer Graphics, 18(12), pp.2917-2926.

[9] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N.A., 2011, June. Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Volume 2 (pp. 42-47).

[10] Hecht, B., Hong, L., Suh, B. and Chi, E.H., 2011, May. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 237-246). ACM.

[11] Schutt, R. and O'Neil, C., 2013. Doing data science: Straight talk from the frontline. O'Reilly Media, Inc.

[12] Keim, D., Kohlhammer, J., Ellis, G. and Mansmann, F., 2010. Mastering the information age solving problems with visual analytics. Eurographics Association.

[13] Massoudi, K., Tsagkias, M., De Rijke, M. and Weerkamp, W., 2011, April. Incorporating query expansion and quality indicators in searching microblog posts. In European Conference on Information Retrieval (pp. 362-367). Springer Berlin Heidelberg.

[14] Albakour, M., Macdonald, C. and Ounis, I., 2013, October. On sparsity and drift for effective real-time filtering in microblogs. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 419-428). ACM.

[15] Makki, R., Soto, A.J., Brooks, S. and Milios, E.E., 2015, December. Active Information Retrieval for Linking Twitter Posts with Political Debates. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on (pp. 238-245). IEEE.

[16] Liu, M., Liu, S., Zhu, X., Liao, Q., Wei, F. and Pan, S., 2016. An uncertainty-aware approach for exploratory microblog retrieval. IEEE transactions on visualization and computer graphics, 22(1), pp.250-259.

[17] Jurgens, D., 2013. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. ICWSM, 13, pp.273-282.

[18] Backstrom, L., Sun, E. and Marlow, C., 2010, April. Find me if you can: improving geographical prediction with social and spatial proximity. In Proceedings of the 19th international conference on WWW (pp. 61-70). ACM.

[19] Davis Jr, C.A., Pappa, G.L., de Oliveira, D.R.R. and de L Arcanjo, F., 2011. Inferring the location of twitter messages based on user relationships. Transactions in GIS, 15(6), pp.735-751.

[20] Peersman, C., Daelemans, W. and Van Vaerenbergh, L., 2011, October. Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

[21] Cheng, Z., Caverlee, J. and Lee, K., 2010, October. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 759-768). ACM.

[22] Ikawa, Y., Enoki, M. and Tatsubori, M., 2012, April. Location inference using microblog messages. In Proceedings of the 21st International Conference on World Wide Web (pp. 687-690). ACM.

[23] Al Zamal, F., Liu, W. and Ruths, D., 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. ICWSM, 270.

[24] Jurgens, D., Finethy, T., McCorriston, J., Xu, Y.T. and Ruths, D., 2015, May. Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In ICWSM (pp. 188-197).

[25] Bakshy, E., Hofman, J.M., Mason, W.A. and Watts, D.J., 2011, February. Identifying influencers on twitter. In Fourth ACM International Conference on Web Seach and Data Mining (WSDM).

[26] Bonsón, E., Torres, L., Royo, S. and Flores, F., 2012. Local e-government 2.0: Social media and corporate transparency in municipalities. Government information quarterly, 29(2), pp.123-132.

[27] Trusov, M., Bodapati, A.V. and Bucklin, R.E., 2010. Determining influential users in internet social networks. Journal of Marketing Research, 47(4), pp.643-658.

[28] Quercia, D., Kosinski, M., Stillwell, D. and Crowcroft, J., 2011, October. Our twitter profiles, our selves: Predicting personality with twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011.

[29] Utz, S., 2010. Show me your friends and I will tell you what type of person you are: How one's profile, number of friends, and type of friends influence impression formation on social network sites. Journal of Computer & Mediated Communication, 15(2), pp.314-335.

[30] LeadSift. 2015. User Insights API. http://leadsift.com/api.html. (2015).

[31] Finkel, J.R., Grenager, T. and Manning, C., 2005, June. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 363-370). Association for Computational Linguistics.

[32] Thomas, J.J. and Cook, K.A., 2006. A visual analytics agenda. IEEE computer graphics and applications, 26(1), pp.10-13.

[33] Ragan, E.D., Endert, A., Sanyal, J. and Chen, J., 2016. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. IEEE transactions on visualization and computer graphics, 22(1), pp.31-40.

[34] Soto, A.J., Kiros, R., Kešelj, V. and Milios, E., 2015. Exploratory Visual Analysis and Interactive Pattern Extraction from Semi-Structured Data. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(3), p.16.

[35] Schreck, T. and Keim, D., 2013. Visual analysis of social media data. Computer, 46(5), pp.68-75.

[36] Manning, C.D., Raghavan, P. and Schütze, H., 2008. Introduction to information retrieval (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.

[37] Baeza-Yates, R., 2016, May. Data and algorithmic bias in the web. In Proceedings of the 8th ACM Conference on Web Science (pp. 1-1). ACM.

[38] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[39] Kenter, T. and de Rijke, M., 2015, October. Short text similarity with word embeddings. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1411-1420). ACM.

[40] O'Connor, B., Krieger, M. and Ahn, D., 2010, May. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In ICWSM (pp. 384-385).

[41] Talburt, J.R., 2013. Overview: The Criticality of Entity Resolution in Data and Information Quality. Journal of Data and Information Quality (JDIQ), 4(2), p.6.