Comparing Methods to Retrieve Tweets: a Sentiment Approach

Stephan Schlosser[1], Daniele Toninelli[2], Michela Cameletti[2]
[1]Center of Methods in Social Sciences, University of Göttingen, Germany, [2]Department of Management, Economics and Quantitative Methods, University of Bergamo, Italy.

Abstract

In current times Internet and social media have become almost unavoidable tools to support research and decision making processes in various fields. Nevertheless, the collection and the use of data retrieved from these sources pose different challenges. In a previous paper we compared the efficiency of three alternative methods used to retrieve geolocated tweets over an entire country (United Kingdom). One method resulted as the best compromise in terms of both the effort needed to set it and the quantity/quality of data collected. In this work we further check, in term of content, whether the three compared methods are able to produce "similar information". In particular, we aim at checking whether there are differences in the level of sentiment estimated using tweets coming from the three methods. In doing so, we take into account both a cross-section and a longitudinal perspective. Our results confirm that our current best option does not show any significant difference in the sentiment, producing scores in between the scores obtained using the two alternative methods. Thus, such a flexible and reliable method can be implemented in the data collection of geolocated tweets in other countries and for other studies based on the sentiment analysis.

Keywords: *social media data collection methods; Twitter data; sentiment analysis; social network; geographical studies.*

1. Introduction

Our society is currently producing an enormous amount of information: just Twitter is able to generate about 500 million of tweets, daily, corresponding to 8TB of data (source: https://www.omnicoreagency.com/twitter-statistics/). These types of big data represent a great opportunity, but also pose several challenges. For example, big data produced using the Internet or social networks can be used in order to support research and decision making processes in several fields. Nevertheless, issues with these data are linked to almost any phase of their "life", starting from the collection phase (e.g. how to collect geolocalized information?) up to their use (new tools are needed to deal with such a huge amount of information), and to their analysis and interpretation (e.g. the representativeness of the covered statistical units). Most of these challenges still need to be fully explored (Goonetilleke *et al*., 2014, Alabdullah *et al.*, 2018 and Morstatter *et al*., 2013).

Our current work is focused on issues linked to the first fase: the collection of social media data. In particular, in a previous work (Schlosser et al., *forth. 2020*) we started studying three alternative methods of collection of messages sent through the Twitter social network (i.e. tweets). These methods were called M1, M2 and M3. The main advantage of all three methods is that they are all able to cover an entire geographical area, United Kingdom (UK) in our case, providing us with a set of fully geolocalized tweets. Nevertheless, the three methods have substantial differences, in terms of level of effort necessary to set them up, of spatial coverage accuracy and of the "amount" of information they are able to produce. Our preliminar study confirmed that, among the three, the best option is M2, a method that reduces the effort to be set (in comparison to M3) and the arbitrarity of decisions of the researcher and problems of overlapping between areas (in comparison to M1). Moreover, M2 produces the same quantity of information (in terms of number of tweets or gigabytes) and enhances the information quality (in terms of number of unique tweets, also reducing the processing times); M2 also leads to a more accurate coverage of the geographical sub-areas studied (UK NUTS; see: https://ec.europa.eu/eurostat/web/nuts/background).

This paper wants to further check if the different settings at the base of the three methods affect the information produced from the content point of view. For this purpose, we analyze tweets collected using all three methods by means of the sentiment analysis applying two of the most widely used lexicons, i.e. AFINN (Nielsen, 2011) and Bing (Bing, 2015). Using such scores, we compare the three methods taking into account both a cross-section (sec. 3.1 and 3.2) and a longitudinal perspective (sec. 3.3). Our expectaction is that there should not be significant differences between tweets collected using the three methods, in terms of level of sentiment (globally and at the sub-area level) and of behavior over time.

Our results confirm these expectations. This further identify M2 as the best option to retrieve geolocalized tweets on a wide geographical area. The high flexibility of such a method allows

to apply it to retrieve geolocalized tweets, setting a certain level of geographical detail and fully covering any other geographical area for any type of research purposes.

2. Literature review

Several recent research projects are based on studies applied to data coming from social networks such as Twitter. These new sources of data, together with the Internet, are also used, from a practical perspective, in order to support decision processes in several fields. In some of these cases researchers require information that is fully geolocalized: this happens, for example, monitoring socio-demographic phenomena (Jashinsky *et al.*, 2014), in disaster management (de Bruijn *et al.*, 2017) or in transportation planning studies (Paule *et al.*, 2019). In this framework, one of the biggest problem is that tweets with a geographical information are just a small fraction of the total (Middleton *et al.*, 2018). Moreover, this information is not always reliable or nicely structured (Middleton *et al.*, 2018, Zheng *et al.*, 2018), mostly because self reported by users. There are currently several methods used to overtake these limits: location extraction (Ozdikis *et al*., 2017, de Bruijn *et al.*, 2017, Zheng *et al.*, 2018) or statistical models and machine learning methods are used to assign spatial coordinates to media items basing on tweets content (Zola *et al.*, 2019, Han *et al.*, 2014). Nevertheless such methods have some limitations, because, for example, the informal and unstructured form of tweets leads to low performances of natural language processing tools (Ajao *et al*., 2015).

Our research aims to overtake such a problem, suggesting one method of tweet collection able to fully cover a geographical area (UK, in our case) and providing tweets that are fully geo-localized. We already compared in a previous work (Schlosser et al., *forth. 2020*) three alternative methods. One of them, M2, resulted the best option, as it is more efficient in comparison to the other two, taking into account the effort for its setting, the "quantity" of information produced as well as the reduction of the data cleaning times. Nevertheless, taking into account the measured sentiment, is M2 also able to perform similarly to the other methods? That is, applying two different sentiment lexicons to tweets collected using all three methods, do we obtain the same sentiment level, distribution and longitudinal evolution?

2. Data and method

In this paper we analyze all tweets collected using our three alternative methods (M1, M2, M3) in the period from January 15 to February 15, 2019. The three methods of collection and their main features are fully introduced in Schlosser *et al*. (2019) and in Schlosser *et al*. (*forthcoming*, 2020). In total, we analyzed 36,348,292 tweets for M1, 40,330,747 for M2 and 34,506,190 form M3, for a total of 111,185,229 tweets.

After first standard steps of cleaning (e.g. removing stop words and special characters and converting the text to lower case), to each tweet collected we apply both the AFINN and the Bing lexicon in order to estimate the level of sentiment. In particular the AFINN lexicon is very widely used for sentiment analysis. Its current version (AFINN-en-165) includes over 3,300 words, each of them associated to an integer score ranging from -5 (very negative) to +5 (very positive). Basing on this, we assign a score to the words of a tweet included in the lexicon and we obtain (summing up such scores) what we define as AFINN score for the considered tweet. The Bing lexicon includes 6,788 words that are classified as positive (we assign to them a value equal to +1) or negative (we assign to them a score equal to -1). The Bing score for a tweet is obtained summing up all the scores linked to the words included in it. Thus, for each tweet we analyzed we obtain two scores (an AFINN and a Bing score).

Our objective is to detect whether there are differences in the level of sentiment score detected using tweets coming from the three different methods. This is done considering three criteria, each of them applied to tweets processed by each of the two lexicons (Bing and AFINN). First, we compare the global averages and the averages by sub-areas (NUTS-1 for UK) computed on all tweets sentiment scores collected by a certain method (sec. 3.1); this is done because the sentiment is a phenomenon strongly varying, at the local level. Second, we compare the distribution of scores (sec. 3.2) to check if the method of collection affects the scores distribution. Third, we check whether our three methods perform similarly in producing sentiment estimates taking into account a longitudinal perspective, i.e. analyzing the evolution of measured sentiment by method and by day (sec. 3.3). This because in studying the sentiment, it is also (or even more) important to detect if the method of collection is able to reproduce accurately the sentiment trend and point-by-point changes over time.

3. Findings

In this section we show the main results of our analysis, according to the three criteria introduced at the end of the previous section.

3.1. Sentiment score comparison

Generally speaking, that is working on all tweets collected using the three studied methods, we did not find any statistically significant difference between the mean scores observed by method. Using the AFINN lexicon, the average score (see last row of Table 1) is equal to 0.769 for M2, to 0.758 for M1 (-1.43%) and to 0.779 for M3 (+2.77%). The average Bing score is equal to 0.247 for M2, to 0.242 for M1 (+2,14%) and to 0.251 for M3 (+3.79%). As a consequence, we confirm that M2 produces results that are intermediate in comparison to the slight underestimation of the sentiment obtained with M2 and the small overestimation obtained with M3. Nevertheless, we applied Kolmogorov-Smirnov tests to verify if there are differences in the distributions of the scores. Both at the level of individual NUTS and at the

country level no significant differences were found among the three methods. By studying the *p*-values we can conclude that the distributions obtained with the three methods are not significantly different (AFINN: M1 vs M2: $p < .001$; M1 vs M3: $p < .001$; M2 vs M3: $p < .001$. Bing: M1 vs M2: $p < .001$; M1 vs M3: $p < .001$; M2 vs M3: $p < .001$)

These very similar results can be caused by the fact that the sentiment observed on such a huge number of tweets and on such a big geographical area (UK) can be driven by a very wide range of topics of different types (pollution, economic scenery, politics, …) that define the current "mood" of Twitter user. Thus, a more reliable analysis was developed working on the Bing and AFINN scores at the level of NUTS (see the first part of Table 1).

	Bing			**AFINN**		
NUTS	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
UKC	0,255	0,246	0,276	0,807	0,771	0,864
UKD	0,226	0,227	0,241	0,708	0,711	0,755
UKE	0,250	0,263	0,244	0,791	0,829	0,777
UKF	0,228	0,226	0,220	0,700	0,702	0,682
UKG	0,251	0,253	0,247	0,764	0,764	0,751
UKH	0,291	0,281	0,289	0,886	0,856	0,885
UKI	0,201	0,185	0,188	0,649	0,598	0,608
UKJ	0,258	0,271	0,273	0,793	0,822	0,827
UKK	0,299	0,300	0,304	0,904	0,907	0,918
UKL	0,281	0,284	0,281	0,852	0,859	0,850
UKM	0,211	0,215	0,209	0,705	0,710	0,686
ALL	**0,242**	**0,247**	**0,251**	**0,758**	**0,769**	**0,779**

Table 1. Average Bing and AFINN scores by NUTS and by method (M1, M2, M3).Table 1 shows that for both lexicons (Bing and AFINN) there are no distributional differences of the scores between the level of individual NUTS (none of the p values is above 0.01).

3.2. Distribution comparison

In Figure 1 we plot the distribution of, respectively, the average Bing and the average AFINN scores obtained analyzing all the tweets by method. Observing both figures it is easy to notice

that there are no differences in the distribution by score computed with the three different methods, both in correspondence to the peaks and along the tail of the distributions.

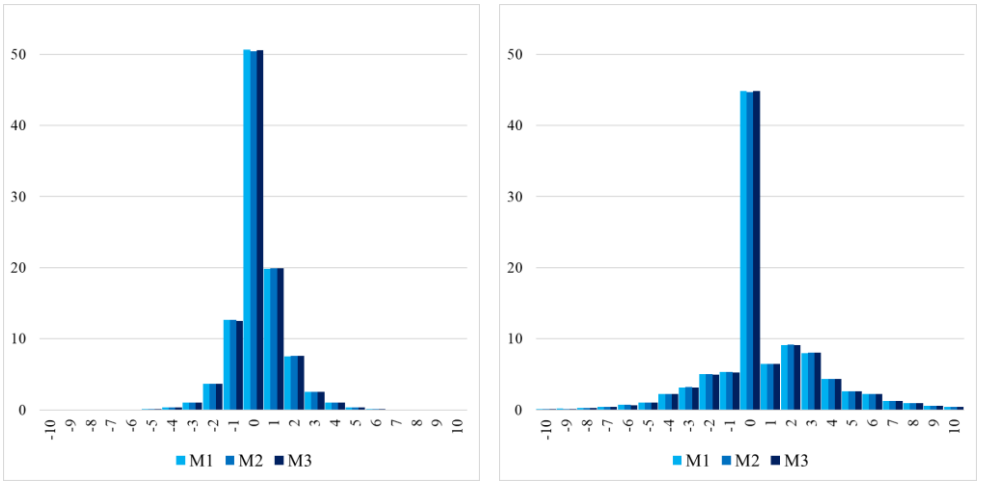

Figure 1. Distribution of the average Bing (left) and AFINN (right) scores by method (all tweets; percentage).

For both distributions we notice high central peaks corresponding to the neutral level. Moreover we notice an higher concentration around the neutral level for Bing lexicon (probably a consequence of the lower variability of the Bing scores assigned to single words) and heavier tails for the AFINN distribution.

3.3. Longitudinal analysis

In order to study potential differences between the three methods of collection in terms of content, it is also relevant to evaluate how the level of sentiment changes over time. This because research can be focused mostly on studying this feature (for a topic, regarding a theme or in a context) rather than in providing a picture referred to a specific time unit. Figure 2, referred to the AFINN lexicon, shows how the level of sentiment changes over the studied month using tweets retrieved by each of the three methods. Graphically, we observe that the three time series are very similar, showing a maximum difference of 4.7% on 2019-02-14 and a minimum difference of 0.7% on 2019-02-11, with a good overlap between the paths of broken lines representing the three methods. The results about the Bing scores are not presented here, because they confirm the findings obtained for the AFINN lexicon (shown in Figure 2). We also computed the correlation between the relative day-by-day changes among the three methods. These correlations are very high and all significantly different from zero (M1 vs M2: $r = .997$, $p < .001$; M1 vs M3: $r = .986$, $p < .001$; M2 vs M3: $r = .989$, $p < .001$).

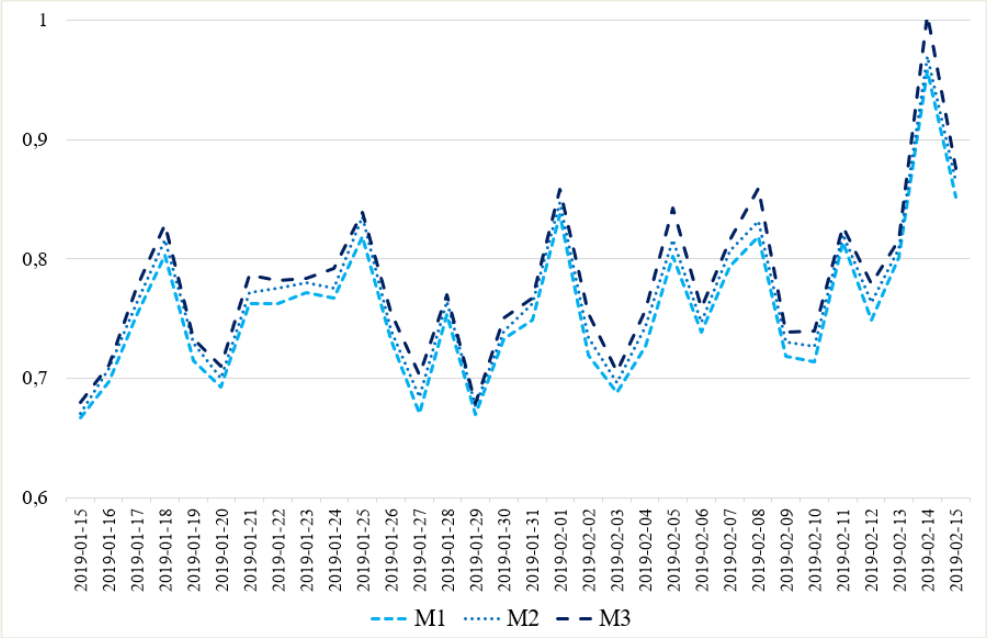

Figure 2. Daily averages of AFINN score by method (Jan. 15 to Feb. 15, 2019).

4. Discussion

In this paper we compare the performances of three methods for retrieving tweets in terms of contents, applying a sentiment analysis by using two of the most common lexicon: AFINN and Bing. Analysing the sentiment at the global level, we notice that M2 produces average scores intermediate between the corresponding scores obtaine using tweet retrieved through M1 and M3. Nevertheless, at the local (NUTS) level this does not hold for all the sub areas. However all the average sentiment scores are not significantly different; thus the three methods can be considered equivalent. If we take a look to the distribution of sentiment scores, we notice two different types of distribution for Bing and AFINN; nevertheless, there are no significant differences between the distributions obtained on tweets retrived by using the three methods. This is further confirmed by the longitudinal analysis (i.e. observing the score daily time series): the relative day by day changes show very small differences between the three methods.

As a final comment, we confirm that M2 is performing very similarly to the alternative methods (and generally produces intermediate results), considering the tweets content. Thus, we can conclude that these results further support our previous findings and that M2, also considering its flexibility, results as the best option in retrieving tweets.

References

Ajao, D., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, 41, 855–864.

Alabdullah, B., Beloff, N., & White, M. (2018). Rise of Big Data - Issues and Challenges. *Proceedings of the 21st Saudi Comput. Soc. Natl. Comput. Conf. NCC 2018*, 0–5.

Bing, L. (2015). Sentiment analysis and opinion mining. New York: Cambridge University Press.

de Bruijn, J., de Moel, H., Jongman, B., Wagemaker, J., & Aerts, J. C. J. H. (2018). TAGGS: Grouping Tweets to Improve Global Geotagging for Disaster Response. *Journal of Geovisualization and Spatial Analysis*, *2*, *2*.

Goonetilleke, O, Sellis, T. K., Zhang, X., & Sathe, S. (2014). Twitter analytics: a big data management perspective. *SIGKDD Explorations*, 16(1), 11-20.

Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.*, 49, 451–500.

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, 35, 51–59.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 400-408.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris Y. (2018). Location Extraction from Social Media. *ACM Trans. Inf. Syst.*, 36(4), article 40.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*, Keraklion, Crete, Greece, 93-98.

Ozdikis, O., Oğuztüzün, H., & Karagoz, P. (2017) A survey on location estimation techniques for events detected in Twitter. *Knowl. Inf. Syst.*, 52(2), 291–339.

Paule, J. D. G., Sun, Y., & Moshfeghi, Y. (2019). On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Inf. Process. Manag.*, 56(3), 1119-1132.

Schlosser, S., Toninelli, D., & Fabris, S. (2019). Looking for Efficient Methods to Collect and Geolocalise Tweets. *Book of Short Papers SIS2019*, Milan, Italy, 1057–1062.

Schlosser, S., Toninelli, D., & Cameletti, M. (*forthcoming*, 2020). Comparing Methods to Collect and Geolocate Tweets.

Zheng, X., Han, J., & Sun, A. (2018). A Survey of Location Prediction on Twitter. *IEEE Trans. Knowl. Data Eng.*, 30, 1652–1671.

Zola, P., Cortez, P., & Carpita, M. (2019). Twitter user geolocation using web country noun searches. *Decis. Support Syst.*, 120, 50–59.