Semiautomatic dictionary-based tweet classification for measuring well-being

M. Cameletti[1], S. Fabris[1,*], S. Schlosser[2] and D. Toninelli[1]

[1] *University of Bergamo (IT); michela.cameletti@unibg.it, silvia.fabris@unibg.it, daniele.toninelli@unibg.it.*
[2] *University of Göttingen (D); stephan.schlosser@sowi.uni-goettingen.de.*
[*] *Corresponding author*

Abstract.

In this paper we describe a semiautomatic dictionary-based approach to filter tweets talking about specific topics. In particular, we are interested in studying the citizen well-being (WB) and, for this aim, we select tweets pertaining two WB dimensions such as environment and health. For this purpose, we use dictionaries containing keywords selected by analyzing tweets published by some Official Social Accounts linked with the two topics. The selected tweets are then processed in order to estimate the sentiment of the population with respect to such specific subjects. In this paper, we present some preliminary results for Great Britain (GB) using tweet collected on the whole country for the six-weeks period from 2019/01/14 to 2019/02/24. The results show that, on the one hand, our dictionary-based classification approach reaches good levels of accuracy, sensitivity and specificity; on the other hand, we assess the spatial variability across GB of the two dimensions we are studying by means of the tweets sentiment analysis.

Keywords. *Twitter; sentiment analysis; health; environment; spatial analysis*

1 Introduction

Measuring individual well-being (WB) is extremely challenging due to the multidimensional, country-specific and latent nature of this concept. Standard approaches for WB evaluation are mainly based on large-scale surveys and rely on several multivariate statistical methods. For example, in [2] the Structural Equation Modelling (SEM) approach is applied to data collected through the European Social Survey (ESS) in order to estimate WB in 16 European countries. In particular, by means of the SEM, the paper identifies seven latent dimensions linked to WB: social involvement, country attachment and trust, discrimination, income perception, environment, health and work status. These dimensions are then used to estimate the WB level of the considered European countries.

Nowadays, in the era of social networks, a huge quantity of data is available that can potentially be used to estimate WB. The collection and the analysis of such data is still an evolving research field that can lead to some advantages: data obtained from the Internet are available at lower costs, in shorter times and are easier to collect than traditional survey-based data. Nevertheless, the collection of this new kind of data is also challenging, from the methodological point of view. Social networks, for example, are used for many different purposes and shared posts can be about personal opinions, ideas, goals and events, but they can also include a huge amount of advertisements and news. For this reason, the identification and

the selection of truly informative data can be a difficult task.

The purpose of this research is to test a reliable semiautomatic dictionary-based method to filter, by topic, posts shared on Twitter, in order to retrieve tweets related to the seven WB dimensions defined in [2]. In particular, in this paper, we focus on two dominions: "health" (HEA) and "environment" (ENV). For this purpose, we define one specific dictionary for each dimension by using a list of keywords chosen by analyzing tweets published by a selected list of Official Social Accounts (OSA).

In this paper we aim to evaluate the reliability of our semiautomatic method using tweets posted in Great Britain (GB). Moreover, since the selected tweets are geolocalized, we are able to study the spatial variability across GB of tweets sentiment, which is a proxy of the two selected WB dimensions.

2 Data and methods

Our data include tweets posted in GB from 2019/01/14 to 2019/02/24 and collected through the "circle approach" described in [1]. Just 1% of these tweets provides GPS coordinates; nevertheless, the "circle approach" allows us to geo-localize all tweets, making it possible to associate each text to one of the circles covering GB (see Figure 1). After having preliminary removed bots (i.e. accounts which post more than 3 times a day, on average), we analyzed 22,193,719 text messages, an average of 26,233 tweets for each circle. In cleaning the corpus of the selected tweets we try to keep as much information as possible by replacing, with equivalent-meaning expressions, htmls, emoji, slangs, word elongations and money symbols; moreover, we keep hashtags and quotations in the tweet text.

The first objective of our analysis was to define two dictionaries to filter among all the available tweets the ones pertaining ENV and HEA. For this purpose, we analyzed several Twitter OSA linked to each of the two dimensions and belonging to no-profit associations, news media, research institutes and intergovernmental organizations[1]. In particular, we collected all the available tweets posted up to 2019/04/04, obtaining 38,604 tweets about ENV and 38,651 about HEA. Our analysis relies on the four following steps.

(1) OSA tweets cleaning. All tweets are cleaned, removing url links, html code, non-ascii and special characters. **(2)** Setting up dictionaries. We select the top trending hashtags used by the selected OSA. These hashtags are keywords used in the OSA description (e.g. #UseLessPlastic for the @LessPlasticUK account) or created by OSA for particular international events (e.g. #PlasticFreeFriday). Among the top trending ones, we selected the most used hashtags: 60 about ENV and 11 for HEA. These thresholds are set, by topic, in order to avoid the selection of acronyms and of too general words (such as for example #women, #plastic, #brexit, #ue, etc.). The selected hashtags constitute the basis of the dictionaries; we then further enrich the list of keywords by analyzing the corpus of the OSA tweets. In particular, we include in the dictionaries the most common bigrams and trigrams (excluding the ones containing stop words). In order to choose combinations of words widely used, we take into account, for each dictionary, bigrams occurred at least 65 times and trigrams occurred at least 35 times. Finally, we manually review the selected hashtags, bigrams and trigrams in order to exclude expressions too generic and not related to the studied WB dimensions (e.g. "facebook live", "fake account", "million people"). The obtained ENV dictionary contains 61 hashtags and 53 bigrams/trigrams; the HEA dictionary includes 11 hashtags and 62 bigrams/trigrams. **(3)** Tweets selection. Using the dictionaries obtained at step (2), among the 22+ millions of tweets collected for GB, we select the ones containing at least one keyword included in the dictionaries. We obtained 35,250 tweets about ENV and 50,610 about HEA. **(4)** Sentiment analysis. These selected tweets are processed by using the AFINN and of the BING lexicon-based approaches. In

[1] List of selected OSA. For ENV: *@climateprogress, @ClimateReality, @friends_earth, @Greenpeace, @GreenpeaceUK, @LessPlasticUK, @PlasticPollutes, @UNEnvironment, @UNFCCC, @World_Wildlife, @WWF, @WWFScotland*. For HEA: *@bbchealth, @CDCgov, @goodhealth, @NBCNewsHealth, @NYTHealth, @EverydayHealth, @NIHClinicalCntr, @theNCI, @CDC_HIVAIDS, @CDCSTD, @CDC_Cancer, @cdchep.*

particular, the first method ranks each word with a score included between -5 and +5 (where negative and positive scores indicate negative and positive sentiment, respectively). The BING lexicon associates -1, 0 and +1 to negative, neutral and positive words, respectively. The total sentiment score of each tweet is computed as the sum of the scores linked to all the words included in the tweet. Thus, for each tweet we obtain two scores (coming from the AFINN and the BING lexicon, respectively).

3 Results

In order to evaluate the performance of our semiautomatic filtering, we selected randomly 100 tweets for each dimension and we manually classify them into two categories: "related" and "non-related" to the topic. To compare the dictionary-based classifications with the manual one (our benchmark), we compute the following performance indexes: *accuracy* (**A**, i.e. the percentage of correctly classified tweets), *sensitivity* (**SE**, i.e. the percentage of topic related tweets correctly identified by the classifier) and *specificity* (**SP**, i.e. the percentage of topic non-related tweets correctly not identified by the classifier). For ENV we obtained the following values for the performance indexes: A=98, SE=97, SP=99. For HEA the observed performance indexes were equal to: A=97.5, SE=95, SP=100. All values denote a very good performance of our semiautomatic dictionary-based approach in filtering tweets related to a given topic.

The sentiment analyses based on AFINN and BING lexicon do not show any significant difference in mean; for this reason (and for space concerns) we present here just the results obtained using BING lexicon. Figure 1 shows the spatial distribution by quartiles of the standardized average sentiment score for ENV (left) and HEA (right). The spatial correlation Moran's index is equal to 0.06 (p=0.04) for ENV and to -0.04 (p=0.90) for HEA, showing absence of relevant spatial correlation between circles. This could be caused by the fact that we are averaging sentiment values across the whole time period and some circles, located in remote zones, may contain a very small number of tweets. Moreover, we implemented a correlation test between sentiment results concerning ENV and HEA, in order to check if the two dimensions influence each other: the correlation is very low and nonsignificant (corr. coeff.=0.09; p=.013).

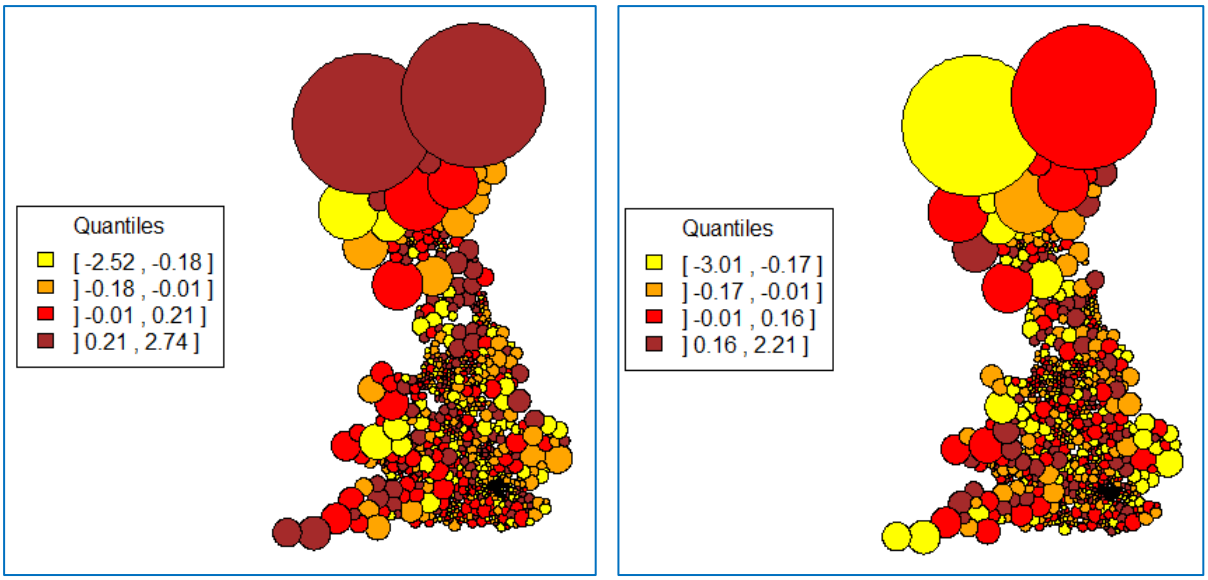

Figure 1: Standardized average tweets sentiment for ENV (left) and HEA (right) (BING lexicon; spatial distribution by quartiles).

4 Conclusions

The aim of this paper is twofold. On the one hand, we want to check if the dictionaries set up by means of our prototypal methodology (see sect. 2) are able to select tweets linked to two WB dimensions, i.e. ENV and HEA. On the other hand, our target is to obtain estimates of the level of two WB dominions over the GB. This second step was based on the sentiment analysis of tweets selected by means of our dictionaries and on the use of the AFINN/BING lexicon-based approaches (the two methods did not show any significant difference in the obtained results).

For what pertains the first objective, all the classification performance indexes (A, SE and SP) show that the both dictionaries perform very well. In fact, they are able to identify posts that have a content actually linked with the dimensions of interest and to exclude the ones which have a content linked to different topics. Our unsupervised topic-classification method is not still fully automatic because we have to select the thresholds, separately for each topic, for the number of hashtags and bigram/trigrams that have to be considered (see Sect. 2) in order to remove keywords not concerning the dimensions of interest. As future research, we intend to include in the dictionary acronyms and to gradually and continuously increment the number of included keywords by periodic analysis of the OSA accounts.

With respect to the second objective, we are able to predict separately for each GB circle the sentiment of the population using tweets about two topics related to WB (ENV and HEA). We expected to find some correlation between neighboring circles and between the two topics within circles. Our findings did not confirm our initial expectations. Moreover, the time lag we took into consideration is probably too short to get a clear picture, that would probably emerge by studying a longer period of analysis (some months of tweets).

Future work will extend the current framework to a different spatial resolution (we will aggregate circles at the area level, by using the so-called NUTS statistical regions of UK); moreover, we will take into account the distribution of sentiment across time, since tweets distribution can be susceptible to daily/weekly events. The final aim of this research is to compare the tweet-based estimation of WB with some benchmark estimates provided by large-scale survey projects, such as the ESS. To this regard, we will consider not only a longer time interval for tweets collection, but also the full set of seven WB dimensions.

References

[1] Schlosser, S., Cameletti, M., Toninelli, D. (2019). Optimized strategies for enhancing the territorial coverage in Twitter data collection. in Keusch, F., Struminskaya, B., Hellwig, O., Stützer, C. M., Thielsch, M., Wachenfeld-Schell, A. (Eds.): 21st General Online Research Conference. Proceedings. Cologne 2019 (link: https://www.gor.de/gor19/index.php?page=browseSessions&form_session=48&presentations=show).

[2] Toninelli, D., Cameletti, M. (2018). Is Structural Equation Modelling Able to Predict Well-being?. In Abbruzzo, A., Brentari, E., Chiodi, M., Piacentino, D. (Eds.). Book of short Papers SIS 2018, 1529-1534, Pearson (link: https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/ISTITUZIONI%20-%20HE%20-%20PDF%20-%20SIS%20V2.pdf).