How to Use AI to Obtain Representative Public Opinion from the Twitter

Erdem Yörük*, Ali Hürriyetoğlu*, Mehmet Fuat Kına*, Fırat Duruşan*, Tolga Etgü*, Osman Mutlu*, Melih Can Yardı*, N. Gizem Bacaksizlar Turbic**, Özgem Elif Acar*, Oğuz Gürerk***, Sukru Atsizelti****

*Koç University, ** GESIS, *** Boğaziçi University, ****Istanbul University

In this paper, we describe our approach in a new research project that aims at scaling up traditional survey polls for public opinion research with AI-based social data analytics. The Politus Project[1] combines econometric and computational methods to create a data platform that delivers representative, reliable, instant, multi-country and multi-language panel data on key political and social trends. The project will collect data from Twitter and process it with deep learning models and natural language processing (NLP) tools that will be developed from the ground up as language-independent and generalizable systems. The platform will deliver geolocated, daily and demographically disaggregated data generated from digital traces of Twitter users on political public opinion, behavior, and attitudes. In terms of public opinion, we produce data on the prevalence of particular political ideologies and values, and approval ratings of political entities based on sentiment, emotion, and stance analyses performed in automatically detected topic categories. We also collect reliable data on demographic characteristics of social media users including income levels, gender, age, ethnicity, race, religion; and political behaviors such as voting and protest participation. Our interdisciplinary approach will be the key to overcoming the complexities of diverse political and societal dynamics in different countries and times by bringing conceptual clarity to the object of analysis, a major challenge in supervised machine learning (ML).

The paper will explain how the Politus Project strives to overcome the following challenges in digital behavioral data, identified by Sen et al (2021): construct validity error, platform affordances error, trace and user selection error, trace and user augmentation error, trace measurement error, platform coverage error and adjustment error.

Construct validity error: The high performance of the automated process is ensured using a gold standard corpus (GSC) as the training, evaluation and test dataset. Following the extensive approach proposed by Hürriyetoğlu et al (2021), a dedicated annotation manual has been prepared for labeling topics, ideologies, emotions, and stance for each country. Each tweet is annotated by at least two experts and disagreements are resolved by an annotation supervisor. High quality is ensured by training the experts on the topics and annotation methodology, regular meetings, and annotation manuals that are updated according to the observations of the experts. This minimizes construct validity error (Sen et al 2021). Overall, the adopted framework was data-driven and linguistically oriented, incorporating a human-in-the-loop approach. The employed methodology is semi-supervised, in the sense that a small fraction of data was labeled and used for the system development. It is linguistically driven, thus morphosyntactic information from basic NLP tools is utilized to identify the information types defined in the Codebook.

Platform affordances error: In order to cope with vocal users as a source of platform affordance error (cf. Jungherr, 2015; Jungherr, Schoen, & Jürgens, 2016), we will make predictions based on users rather than their posts when inferring about social phenomena. Once our classification models predict certain characteristics (e.g.,being conservative), we will identify the probability of users to have such characteristics in a time-variant manner in order to have a better out-of-platform representativeness.

Empirical test: Comparison between content-based and user-based predictions.

Trace and user selection error: The first step is to set up policies and software to continuously collect, acquire, and organize data (Sang and van den Bosch 2013, Steinberger 2013, Hürriyetoğlu 2016). Creation

[1] The Politus Project has been funded by the ERC Proof of Concept and by the Scientific and Technological Research Council of Turkey (TÜBİTAK) and it will extend the technological scope of the ERC-funded Emerging Welfare Project (emw.ku.edu.tr).

of unbiased random samples, development of pre-trained deep learning models (Grurangan et al. 2020) and detecting spatio-temporal trends depend on the availability of such data. Users posting from a certain case country, Turkey initially, are collected by utilizing the most followed users in this particular country, rather than an API-based search strategy. Our dataset on Turkey contains all 53 million unique users (since the launch of Twitter in Turkey). Access to all users, increases representativity of this data, in the first place. Then we have identified 3 million users whose gender and location are determined from their Twitter meta data. Finally, we collect the most recent posts, likes, followers, followees, quotes, retweets, and replies of all users in our dataset. On the other hand, the representativeness of our dataset over the Twitter space is also an important issue that emerges in the ML process. In order to create the annotated corpus and to choose the tweets to annotate, we rely on random sampling rather than keyword filtering (Yörük et al 2021). While this strategy increases the cost of annotation, it maximizes the capacity for generalization. Since the automatic classification and aggregation of data will depend on NLP tools that are trained on a human-annotated corpus, the representation of constructs in signals will not be dependent on existence of keywords or other features that have limited applicability. Since annotators will code random samples of tweets, they will likely encounter diverse expressions that display the signals, securing the representative and comprehensive nature of the signals. This is also expected to diminish, if not eliminate, the issue of user selection error since an annotated model is less likely to be biased against categories of users compared to restrictive keyword lists or search engines.

Empirical test: Comparison between API-based and follower-based user collection strategy and comparison between keyword-based and random sampling of GSC

Trace and user augmentation error: The project creates a large (24,970 tweets as of June 2022), high quality, human-annotated GSC from randomly sampled social media data to train, test and evaluate the NLP tools to ensure accuracy and completeness of the data they generate. We will evaluate the validity of automatic data both internally, seeking a minimum F1 score of 0.75, and externally, seeking a minimum of 0.9 correlation with external ground-truth indicators, such as official statistics or reliable surveys (following Blumenstock, J., Cadamuro, G., & On, R. (2015)). Topic analysis is applied to the text collections to identify the most salient features to classify social trends (Park and Lee 2020, Sukhija 2016). Next, sentiment, emotion, and stance analysis enable understanding the public opinion (Li and Caragea 2019). Network analysis and information diffusion on social media data shed light on the reach and bias of the information (Taxidou and Fischer 2013, Benoit 2020). Finally, linking and combining different datasets contribute to deriving unique and significant insights (Khaefi et al. 2018, Pulse UN Global 2016, Theocharis and Jungherr 2021).

Empirical test: Validating with ground-truth data.

User reduction error: Automatic, irrelevant and malicious Twitter accounts, such as bots, that may affect the political opinion measurement process are detected and removed from the dataset using botometers.

Descriptive statistics of irrelevant users

Trace measurement error: ML-based modeling will be the main AI approach to generate the dataset. The data infrastructure will utilize the deep learning models in many languages, which are Multilingual Bidirectional Encoder Representations from Transformers (mBERT) (Devlin et al., 2019), and Cross-lingual Language Model (XLM) (Conneau et al., 2020). For the ML models, two different types of training data will be used: The first is the GSC of annotated tweets. For each task, the GSC is divided into three subsets for training, validation and evaluation and using these, multilingual deep learning-based ML models are fine-tuned. By analysing the contents posted by users, the subjects are determined for each geographical unit, and emotional and stance profiles related to them are created. By adding the data shared in the past, this process generates the temporal distribution in real-time for each geographical region. The models depend on ML-based sentiment and stance analysis, besides network-based measures, such as structural and centrality metrics. Second, we will follow in the footsteps of Blumenstock et al. (2015) and connect survey data and social media data. We will conduct an online survey at Facebook, where the

questionnaire will include questions about demographical, political and sociological. Then, we will collect tweets of the participants of this survey with their consent and build ML models that will predict these characteristics collected in the survey using the content of tweets. Then, we will use this model to predict the characteristics of all Twitter users in our dataset. We will repeat these online surveys every month in order to keep our models updated, essentially because of the time-variant nature of the indicator that we are striving to extract. Also, the primary function of the GSC is the evaluation of zero-shot cross-lingual deep learning models that are based on mBERT and XLM-R and perform well under low-resource settings. Active learning is also applied to detect the documents that should be annotated to increase the efficiency of the annotation and ensure high performance of the text processing tasks

We use network analysis to improve the performance of our determination of demographic and political characteristics of users. The Politus data is collected in both user- and network-oriented manners, which reveals the location, demographic, and socio-economic characteristics of the users and helps investigate political homophily in the ties of users (Colleoni et al. 2014,Wagner 2017). The analysis of interaction networks between Twitter users might be more dependent on the data collection design than the analysis of the message content as the random missing data can affect the number of users and the interactions in social networks (Jungherr 2016). Therefore, we collect tweets for the analyzed periods and compare different datasets and network measures to investigate results from user and tweet databases.

Empirical tests: Comparison between annotation-based ML and Facebook-survey-based ML and comparison between trace measurement with or without network analysis

Platform coverage and adjustment error: Using the data extracted from online networks for public opinion is often criticized as having a solid demographic bias - participation in online networks is strongly affected by users' age, sex, education level, race, income level, etc. (Mislove et al. 2011; Sloan 2017; An and Weber 2015; Cohen and Ruths 2013; Filho et al. 2015; Barbera 2016; Olteanu et al. 2019; Sen et al. 2021). Our Twitter data is non-probabilistic and it is challenging to generalize our findings to the real population (Sen et al 2021). To deal with demographic bias and overcome the non-probability sample characteristics of the data, we will utilize Bayesian multilevel regression with poststratification (MRP). Multilevel regression provides an estimate for each group with a common combination of population characteristics with high precision even for those with relatively few representatives in the data while poststratification provides an adjustment to remove the aforementioned bias by assigning a weight to each group according to its actual representation in the general public (Lax and Phillips 2009; Park, Gelman and Bafumi 2004; Gelman 2007; Gelman and Little 1997). The state-of-the-art Bayesian version of the model will be used by focusing on the key characteristics of age, gender, level of education, and location.

Empirical tests: Comparison between models with and without MRP.