



# All Azimuth

ISSN: 2146-7757 (Print)

---

## **Automated Text Analysis and International Relations: The Introduction and Application of a Novel Technique for Twitter**

Emre Hatipođlu, Osman Zeki Gökçe, İnanç Arın and Yücel Saygın

To cite this article: Emre Hatipođlu, Osman Zeki Gökçe, İnanç Arın and Yücel Saygın, "Automated Text Analysis and International Relations: The Introduction and Application of a Novel Technique for Twitter," *All Azimuth* 0, no. 0 (2018): 1-22, doi: 10.20991/allazimuth.476852.

Published Online: 27 December 2018

## Automated Text Analysis and International Relations: The Introduction and Application of a Novel Technique for Twitter

Emre Hatipoğlu, Osman Zeki Gökçe\*, İnanç Arın and Yücel Saygın  
Sabancı University

\* İstanbul Medipol University

---

### Abstract

*Social media platforms, thanks to their inherent nature of quick and far-reaching dissemination of information, have gradually supplanted the conventional media and become the new loci of political communication. These platforms not only ease and expedite communication among crowds, but also provide researchers huge and easily accessible information. This huge information pool, if it is processed with a systematic analysis, can be a fruitful data source for researchers. Systematic analysis of data from social media, however, poses various challenges for political analysis. Significant advances in automated textual analysis have tried to address such challenges of social media data. This paper introduces one such novel technique to assist researchers doing textual analysis on Twitter: More specifically, we develop a clustering methodology based on Longest Common Subsequence Similarity Metric, which automatically groups tweets with similar content. To illustrate the usefulness of this technique, we present some of our findings from a project we conducted on Turkish sentiments on Twitter towards Syrian refugees.*

**Keywords:** Social media, foreign policy, refugee, public opinion, Turkey

### 1. Introduction

In many polities around the globe, quick and far-reaching dissemination of information on social media have been redefining politics. In regimes based on popular support, politicians and policy makers have increasingly become sensitive and responsive to online perceptions and attitudes of the public. Social media platforms such as Twitter, Facebook, Pinterest etc. are becoming the new loci of political communication, debate, propaganda and mobilization, increasingly circumventing traditional channels for politics.

Accordingly, various interrelated research strands have emerged on social media and in international politics. Scholars have aptly illustrated how social media have transformed various basic concepts such as diplomacy,<sup>1</sup> warfare,<sup>2</sup> transnational activism,<sup>3</sup> and political

---

Emre Hatipoğlu, Faculty of Arts and Social Sciences, Sabancı University. Email: ehatipoglu@sabanciuniv.edu. Osman Zeki Gökçe, Faculty of Humanities and Social Sciences, İstanbul Medipol University. Email: ogokce@medipol.edu.tr. İnanç Arın, Department of Computer Science and Engineering, Sabancı University. Email: inanc@sabanciuniv.edu. Yücel Saygın, Faculty of Engineering and Natural Sciences, Sabancı University. Email: ysaygin@sabanciuniv.edu.

<sup>1</sup> Philip Seib, *Real-Time Diplomacy: Politics and Power In the Social Media Era* (New York, NY: Palgrave Macmillan US, 2012); Nicholas J. Cull, "The Long Road to Public Diplomacy 2.0: The Internet in US Public Diplomacy," *International Studies Review* 15, no. 1 (2013): 123–39; Jan Melissen and Emillie V. de Keulenaar, "Critical Digital Diplomacy as a Global Challenge: The South Korean Experience," *Global Policy* 8, no. 3 (2017): 294–302.

<sup>2</sup> Holger Pötzsch, "The Emergence of iWar: Changing Practices and Perceptions of Military Engagement in A Digital Era," *New Media & Society* 17, no. 1 (2015): 78–95.

<sup>3</sup> Wilma de Jong, Martin Shaw, and Neil Stammers, eds., *Global Activism, Global Media* (London, UK: Pluto Press, 2005).

development.<sup>4</sup> The literature on social media is burgeoning in Turkey as well. Recent studies have examined the relationship between social media and social movements,<sup>5</sup> who the opinion leaders are on Turkish Twitter,<sup>6</sup> and to what extent online opinion clusters on Twitter map onto conventional public opinion clusters in Turkish society.<sup>7</sup> A wealth of studies on social media and Turkey also exist in other areas of social sciences such as marketing<sup>8</sup> and public health.<sup>9</sup>

Systematic analysis of data from social media poses various challenges for political analysis. Most existing approaches, while offering excellent tools for conventional text-data sources such as news reports and international treaties, fail to address this dynamic nature of social media. Social media is virtually open to the globe; its crowd-sourced structure creates data in very large amounts. The language used is dependent on the context, does not necessarily follow a preset syntax and transforms very quickly. Similarly, topics of interest emerge and evolve at great speed, making spontaneous analysis almost a necessity.

Significant developments in automated textual analysis have tried to address such challenges of social media data. These techniques aim to do away with manual labor needed to sort and code data in extremely large quantities. This paper introduces one such novel technique to assist researchers doing textual analysis on Twitter. The technique develops a measure based on Longest Common Subsequence Similarity Metric (LCSSM), which automatically clusters tweets with similar content. More specifically, the technique decouples meaning from character in text, and develops a measure of character similarity across strings of text. This technique significantly reduces the workload of manual coding for content, this reduction in workload usually ranges from 75% to 90%. While LCSSM does not necessarily do away with the need for human coding, it presents an easy-to-use and scalable tool to conduct sentiment analysis on Twitter. To illustrate the usefulness of this technique, we present some of our findings from a project we conducted on Turkish sentiments towards Syrian refugees on Twitter.

This manuscript consists of five sections. The next section will give an overview of automated text analysis in international relations literature, and section two will introduce basic terminology and main techniques utilized in automated text analysis. Section three will describe why the LCS similarity metric was developed, and how this technique fares vis-à-vis existing techniques of automated text analysis. In doing so, it will also present a discussion on what types of fundamental differences data from Twitter exhibit compared to texts traditionally employed in international relations research. This section also includes an

<sup>4</sup> Shanthi Kalathil and Taylor C. Boas, *Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule* (Washington, DC: Carnegie Endowment, 2010).

<sup>5</sup> See for example Zeynep Tüfekçi, *Twitter and Tear Gas: The Power and Fragility of Networked Protest* (New Haven, CT: Yale University Press, 2017); Şenay Yavuz Görkem, "The Weakest Link or the Magic Stick?: Turkish Activists' Perceptions on the Scope and Strength of Digital Activism," *Turkish Studies* 18, no. 1 (2017): 102–24; H. Akın Ünver and Hassan Alassaad, "How Turks Mobilized Against the Coup: The Power of the Mosque and the Hashtag," *Foreign Affairs*, September 14, 2016; Kamil Demirhan, "Social Media Effects on the Gezi Park Movement in Turkey: Politics under Hashtags," in *Social Media in Politics: Case Studies on the Political Power of Social Media*, ed. Bogdan Pătruț and Monica Pătruț (Basel: Springer International Publishing, 2014), 281–314.

<sup>6</sup> Osman Z. Gökçe, Emre Hatipoğlu, Gökhan Göktürk, Brooke Luetgert, and Yücel Saygın, "Twitter and Politics: Identifying Turkish Opinion Leaders in New Social Media," *Turkish Studies* 15, no. 4 (2014): 671–88.

<sup>7</sup> Emre Hatipoğlu, Osman Z. Gökçe, Berkay Dinçer, and Yücel Saygın, "Sosyal medya ve Türk dış politikası: Kobani tweetleri üzerinden Türk dış politikası algısı." *International Relations/Uluslararası İlişkiler* 13, no. 52 (2016): 175–97.

<sup>8</sup> See for example İrem Eren Erdoğan and Mesut Cicek, "The Impact of Social Media Marketing on Brand Loyalty," *Procedia-Social and Behavioral Sciences* 58 (2012): 1353–60.

<sup>9</sup> See for example Christopher H. Purdy, "Using the Internet and Social Media to Promote Condom Use in Turkey," *Reproductive health matters* 19, no. 37 (2011): 157–65.

account of the research teams' personal experience with automated text analysis techniques, and the available training options for an IR scholar to be proficient in such techniques. Section four presents an application of the LCSSM technique conducted on Turkish foreign policy where the authors assess the relative salience of policy issues pertaining to Syrian refugees in the tweets of Turkish users. The final section concludes with a discussion on further applications of the LCSSM technique in IR and beyond.

## 2. Automated Text Analysis and International Relations

Automated text analysis opens many possibilities for international relations scholars. Automated text analysis is scalable; since storage and computing power is virtually free, the same program can work with small or very large datasets. As a source of data, text offers various types of content that are of use to academics, such as categories and sentiments. Among these by-products, "assigning texts to categories is the most common use of content analysis methods in political science," in general, and international relations, in particular.<sup>10</sup> These categories are often pre-defined by the research question at hand, and the literature this research question builds on. Computers are, then, taught to allocate each unit of analysis under these pre-defined categories. Categorizations can take on a binary (e.g. yes/no: occurred/did not occur), thermometer (e.g. strongly against-against-neutral-in favor-strongly in favor), or nominal (e.g. Africa-America-Asia-Europe-Oceania) nature. The derivation of such mutually exclusive categories also allow the positioning of texts (or the individuals that produce them) in multidimensional platforms, hence allowing spatial analysis of political interaction.

Three main approaches exist for categorization of text using computers, namely nonstatistical dictionary based (DB), supervised machine learning (SML) and unsupervised machine learning (UML).<sup>11</sup> DB is arguably the simplest and the oldest of the three. In this technique, the categories are pre-defined with a pre-set dictionary including all the keywords associated with these categories. Creating these dictionaries constitute the most time and labor-intensive part of automated text analysis techniques. Given a dictionary, the computer takes a tally of key words in a text, compares the prevalence of these key words against this pre-set dictionary and assigns texts to categories. For instance, Adler and Wilkinson allocate each piece of Congress bill to a predetermined issue area such as education, environment or defense.<sup>12</sup> In doing so, they use a pre-built dictionary. Project Civil Strife<sup>13</sup> collects various reports on state, insurgent and civilian actions in Indochina and allocates each report to a conflict event using a pre-set dictionary.

The exponential increase in computing power have allowed the development statistical classification techniques in which computers derive categories from existing text. These statistical-based methods can be grouped under two categories: supervised machine learning (SML) and unsupervised machine learning (UML). Supervised machine learning does not identify a list of key words per se. Instead, researchers identify "reference" texts that approximate the ideal categories and feed these reference texts to the computer program. To

<sup>10</sup> Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise And Pitfalls Of Automatic Content Analysis Methods For Political Texts," *Political Analysis* 21, no. 3 (2013): 267–97, 272.

<sup>11</sup> Lori Young and Stuart Soroka, "Affective News: The Automated Coding of Sentiment in Political Texts," *Political Communication* 29, no. 2 (2012): 205–31.

<sup>12</sup> E. Scott Adler and John D. Wilkerson, *Congress and the Politics of Problem Solving* (Cambridge, UK: Cambridge University Press, 2013).

<sup>13</sup> Stephen M. Shellman, "Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors over Time and Space," *Political Analysis* 16, no. 4 (2008): 464–77.

illustrate, party manifestos of socialist parties or the speeches of the leaders of the parties can “teach” a computer program what content can be classified as left-leaning.<sup>14</sup> *Wordscores*<sup>15</sup> is a readily available program that can derive multiple dimensions given a set of existing reference texts such as those found for political parties.

Of the many projects that use automated text analysis techniques used international relations (IR) studies, two require special mention, namely the Computer Events Data System (CEDS) and the Militarized Interstate Disputes (MID) dataset. Despite having originated from different strands of IR literature, both projects categorize various international events. Towards this end, both projects have used SML content analysis, and developed unique dictionaries and training sets to “teach” their respective programs how to filter and categorize content appropriately.

The CEDS (previously called the Kansas Event Data System) has been an ongoing project since the early 1990s and provides one of the most important datasets for positive foreign policy analysis. The CEDS evaluates news reports from established sources such as Reuters and Lexis/Nexis to see (i) whether a given report identifies an international event that occurred between two countries, and, if so, (ii) what specific event took place between these two countries.<sup>16</sup> The CEDS uses the World Event Interaction Survey (WEIS) classification.<sup>17</sup> WEIS identifies several dozen distinct diplomatic events, ranging from extending foreign aid to military engagement as adversaries. For such evaluations, the CEDS operates on a well-developed dictionary that supervises the machine’s learning on whether a report indicates an event or not. Later studies based on the CEDS have modified this dictionary towards research questions other than diplomatic events such as international mediation<sup>18</sup> or civil war.<sup>19</sup>

Another recent application of SML technique was the updating of the Militarized Interstate Disputes (MID) dataset. The MID dataset, which codes all instances of militarized interstate disputes for all country-pairs (dyads) from 1815 until 2010, is one of the most widely used datasets in quantitative studies of conflict.<sup>20</sup> The dataset identifies 22 separate militarized incidents that can take place between two countries, such as verbal threats to use force, mobilization, border fortification, seizure of property, joining an interstate and the use of nuclear weapons, among others.<sup>21</sup> These individual incidents are then clustered under a nominal scale of dispute intensity: no action, threat to use force, display of force, use of force

<sup>14</sup> For an application of this idea to Danish politics, see Robert Klemmensen, Sara Binzer Hobolt, and Martin Ejnar Hansen, “Estimating Policy Positions Using Political Texts: An evaluation of the Wordscores approach,” *Electoral Studies* 26, no. 4 (2007): 746–55.

<sup>15</sup> Michael Laver and Kenneth Benoit, “Locating TDs in Policy Spaces: The Computational Text Analysis of Dáil Speeches,” *Irish Political Studies* 17, no. 1 (2002): 59–73.; Michael Laver, Kenneth Benoit, and John Garry, “Extracting Policy Positions from Political Texts Using Words as Data,” *American Political Science Review* 97, no. 2 (2003): 311–31.

<sup>16</sup> Philip A. Schrodt and Deborah J. Gerner, “Validity Assessment of A Machine-Coded Event Data Set for the Middle East, 1982–92,” *American Journal of Political Science* (1994): 825–54; Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle, “Political Science: KEDS-A Program For The Machine Coding of Event Data,” *Social Science Computer Review* 12, no. 4 (1994): 561–87.

<sup>17</sup> Charles A. McClelland, Rodney G. Tomlinson, Ronald G. Sherwin, Gary A. Hill, and Herbert L. Calhoun, *The Management and Analysis of International Event Data: A Computerized System for Monitoring and Projecting Event Flows* (University of Southern California Los Angeles School of International Relations, 1971); Joshua S. Goldstein, “A Conflict-Cooperation Scale for Weis Events Data,” *Journal of Conflict Resolution* 36, no. 2 (1992): 369–85.

<sup>18</sup> Deborah J. Gerner, Philip A. Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr, “Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions” (paper presented at the Annual Meeting of the International Studies Association, New Orleans, March 2002).

<sup>19</sup> Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor, “Integrated data for events analysis (IDEA): An event typology for automated events data development,” *Journal of Peace Research* 40, no. 6 (2003): 733–745.

<sup>20</sup> Glenn Palmer, Vito d’Orazio, Michael Kenwick, and Matthew Lane, “The MID4 Dataset, 2002-2010: Procedures, Coding Rules and Description,” *Conflict Management and Peace Science* 32, no. 2 (2015): 222–42.

<sup>21</sup> Michael R. Kenwick, Matthew Lane, Benjamin Ostick, and Glenn Palmer, “Codebook for the Militarized Interstate Incident Data, version 4.0” (unpublished manuscript, 2013).

and war. Traditionally, MID data were collected manually; humans searched for relevant news from sources such as New York Times, Keesing's World News Almanac, Facts on File, and later, with the advent of online sources, Factiva/Reuters and Lexis/Nexis.<sup>22</sup> The news collected were, then, manually coded. Having humans conduct relevant searches of news sources had made updating the dataset in a timely manner a very costly and time-consuming endeavor. Instead, the MID team chose to employ support vector machine document classification algorithm to "identify news stories that contain codable militarized interstate actions," and hence, "replacing the most costly and labor-intensive part of the MID coding process."<sup>23</sup> Later advances have allowed these news sources to be coded automatically as well.<sup>24</sup> During this process, the computer algorithm was fed substantial amount of news reports, which the researchers knew contained a specific militarized incident between two countries. Projects that use similar techniques, but are unrelated to the MID group, include the Global Terrorism Database<sup>25</sup> and the GATE (government actions in terrorism environments) project.<sup>26</sup>

More recent techniques do away with human supervision altogether, and let the data determine what categories exist within.<sup>27</sup> Such techniques do not calculate the "similarity" of each unit of text against a set of pre-defined parameters. Rather, they identify the clusters these units form on a multi-dimensional space. For instance, Grimmer and King developed a UML program that has the capacity to cluster data in dimensions initially inaccessible to the human eye.<sup>28</sup> Similarly, Hopkins and King adopt a non-parametric data-driven approach to let the computer derive and calculate proportions of text data, which does not rest on any "parametric statistical modeling, individual document classification or random sampling."<sup>29</sup> The program they developed for R platform, README, is readily available online. Analyzing Twitter data in the aftermath of the US financial meltdown, Bollen, Mao, and Pepe's algorithm derives six distinct categories of sentiments (tension, depression, anger, vigor, fatigue, confusion) without any prior conditioning by researchers.<sup>30</sup> *Wordfish* is another accessible program for those who want to start exploring UML techniques in textual analysis.<sup>31</sup> Finally, this fully-data driven approach has started informing how we collect data at the outset as well. King, Lam and Roberts, for example, take issue with which search words we use to collect data start with, and instead let computers iterate search keywords to derive more relevant data.<sup>32</sup>

<sup>22</sup> Faten Ghosn, Glenn Palmer, and Stuart A. Bremer, "The MID3 Data Set, 1993-2001: Procedures, Coding Rules, and Description," *Conflict Management and Peace Science* 21, no. 2 (2004): 133-54.

<sup>23</sup> Philip A. Schrodt, Glenn Palmer, and Mehmet Emre Hatipoğlu, "Automated Detection of Reports of Militarized Interstate Disputes Using the SVM Document Classification Algorithm" (paper presented at the Annual Meeting of the American Political Science Association, Boston, 2008).

<sup>24</sup> Palmer, et al., "The MID4 Dataset".

<sup>25</sup> Gary LaFree, Laura Dugan, and Erin Miller, *Putting Terrorism in Context: Lessons from the Global Terrorism Database* (London, UK: Routledge, 2014).

<sup>26</sup> Laura Dugan and Erica Chenoweth, "Government Actions in Terror Environments (GATE): A Methodology that Reveals How Governments Behave Toward Terrorists and Their Constituencies," in *Handbook of Computational Approaches to Counterterrorism*, ed. V. S. Subrahmanian (New York, NY: Springer, 2013), 465-86.

<sup>27</sup> Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn, "Fightin'words: Lexical Feature Selection and Evaluation for Identifying The Content of Political Conflict," *Political Analysis* 16, no. 4 (2008): 372-403.

<sup>28</sup> Justin Grimmer and Gary King, "General Purpose Computer-Assisted Clustering and Conceptualization," *Proceedings of the National Academy of Sciences* 108, no. 7 (2011): 2643-50.

<sup>29</sup> Daniel J. Hopkins and Gary King, "Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability," *Public Opinion Quarterly* 74, no. 2 (2010): 201-222, 236.

<sup>30</sup> Johan Bollen, Huina Mao, and Alberto Pepe, "Modeling Public Mood and Emotion: Twitter sentiment and socio-economic phenomena," *ICWSM* 11 (2011): 450-453.

<sup>31</sup> Sven-Oliver Proksch and Jonathan B. Slapin, "WORDFISH: Scaling software for estimating political positions from texts," *Version 1* (2008): 323-344. For a comparison of validity between Wordfish and Wordscores programs, see Frederik Hjørth, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard, "Computers, Coders, and Voters: Comparing automated methods for estimating party positions," *Research & Politics* 2, no. 2 (2015): 1-9.

<sup>32</sup> Gary King, Patrick Lam, and Margaret E. Roberts, "Computer-Assisted Keyword and Document Set Discovery from

### 3. A Novel Method: The Longest Common Subsequence Metric and Clustering Tweets

Our clustering algorithm based on the Longest Common Subsequence similarity metric was developed under the auspices of the *Identifying Political Opinion Shapers on Twitter in Turkey (I-POST)* project. The project has been funded by an internal research grant by Sabancı University, and is the result of three professors, the initial co-principal investigators, coming from three different angles to automated textual analysis.<sup>33</sup> Dr. Emre Hatipoğlu worked from 2007-2010 as a research assistant in the MID project that aimed to transfer data collection from hand coding to automated coding. In doing so, his research team systematically measured the accuracy of computer assignments of categories using support vector machines.<sup>34</sup> Dr. Yücel Saygın, a computer scientist, has been a leading figure on big data and data privacy. Further application of his work has led him to develop modules analysis of social media data for marketing research and product development. Dr. Luetgert's previous research utilized automatized data collection and analysis techniques to focus on how the European *acquis* was transposed to member state legislation.<sup>35</sup> Of the other two co-authors of this manuscript, Dr. Arın (2017) wrote his dissertation on "*Impact Assessment and Prediction of Tweets and Topics*,"<sup>36</sup> and Dr. Gökçe has been the head assistant of the I-POST project; his previous work looked at who the opinion leaders were on Turkish twitter.<sup>37</sup>

Both Dr. Hatipoğlu and Dr. Gökçe have worked with datasets throughout their academic lives. For his dissertation, Dr. Gökçe constructed a global dyadic dataset of interstate energy trade and interdependence which covers the years 1978-2012. The decision for these scholars to work with datasets is related to the inferential leverage that large-N studies provide for certain research questions, in particular, questions that focus on establishing trends and correlations in IR. The canonical debate on whether balance of power leads to peaceful relations in the international system illustrates this leverage. In *Diplomacy*, one of the most important IR works based on realism, Kissinger argues that the stable balance of power system established by the European pentarchy (Austria-Hungary, Britain, France, Prussia and Russia) led to peaceful relations as long as these countries were free to align themselves as they wished.<sup>38</sup> To prove his point, Kissinger renders a lengthy treatise of 19<sup>th</sup> century European diplomatic history. While his account of 19<sup>th</sup> century Europe happens to be quite in line with the assertion that balance of power leads to peace, Kissinger fails to look at cases of (i) balance of power and no peace, and (ii) imbalance of power and peace. The nature of our inferences changes when we look at the whole globe over since the Congress of Vienna in 1815. A wealth of analyses that use large datasets and look at all possible combinations of balance (imbalance) of power and peace (conflict) have consistently shown that balance of power increases the propensity for conflict between states whereas power preponderance leads to more cordial relations.<sup>39</sup>

---

Unstructured Text," *American Journal of Political Science* 61, no. 4 (2017): 971–88.

<sup>33</sup> The third co-principal investigator, Dr. Brooke Luetgert is no longer at Sabancı University. She is not involved with the research outputs presented in this study.

<sup>34</sup> Schrodt, et al., "Automated detection".

<sup>35</sup> Thomas König, Brooke Luetgert, and Tanja Dannwolf, "Quantifying European Legislative Research: Using CELEX and PreLex in EU legislative studies," *European Union Politics* 7, no. 4 (2006): 553–74.

<sup>36</sup> İnanç Arın, *Impact Assessment & Prediction of Tweets and Topics* (Ph.D. thesis, Sabancı University, 2017).

<sup>37</sup> Gökçe, et al., "Twitter and Politics".

<sup>38</sup> Henry Kissinger, *Diplomacy* (New York, NY: Simon and Schuster, 1994).

<sup>39</sup> See for example A.F.K. Organski and Jacek Kugler, *The War Ledger* (London, UK: University of Chicago Press, 1981); Melvin Small and Joel David Singer, *Resort to Arms: International and Civil Wars, 1816–1980* (London, UK: Sage Publications, Inc, 1982); Alex Braithwaite and Douglas Lemke, "Unpacking Escalation," *Conflict Management and Peace Science* 28, no. 2 (2011): 111–23.

Using comprehensive datasets, hence our focus on big data, is equally important for studies that focus on discourse. An often-witnessed inferential pitfall in studies that use small-N to derive correlations and trends in discourse, relates to selective use of evidence to support an argument without due regard to possible alternatives. In such practice, researchers build a theoretical construct, then select events/cases that support this construct while ignoring the possible cases where their theoretical assertions could be falsified. Let us illustrate this faulty practice within our topic of research. Some may argue that the tone in Turkish twitter, in general, is quite unforgiving and critical of Syrian refugees. Furthermore, these scholars could also argue that economic reasons undergird such an offensive tone. To support this argument, then, one can search Twitter history and come up with hundreds of anti-refugee tweets, many of them bordering on outright racism, which also happen to mention the economic burden these refugees bring. Such an inference, however, would be faulty. We would not know the ratio of these anti-refugee tweets to the pro- and neutral-refugee tweets. Likewise, absent a comprehensive and representative dataset, we would not be able to assess the relative salience of economics as a determinant of anti-refugee discourse. As a matter of fact, our results presented below indicate that tweets that are concerned with the refugees' own security and well-being occupy a larger share in Turkish online discussion space than tweets criticizing the economic and social burden the refugee inflow have imposed on Turkey. Further analyses also indicate that, even in such critical tweets, most of the blame is laid on the government, and not the refugees themselves.

### 3.1. The choice of methodology: an effective and efficient tradeoff

Our main methodological challenge, i.e. coding large amounts of unstructured text data led us to adopt a hybrid design, integrating automated clustering with manual coding of content. Our methodology strikes a balance between the reduction in workload that automated content and the accuracy of coding (content validity) manual coding provides. The nature of Twitter data has necessitated such a balance. Doing so came at a cost; not fully automating coding of content came at the expense of a sizeable number of person-hours. Likewise, clustering tweets inevitably led to a small compromise in measurement validity as each cluster header speaks for all tweets under its cluster. Some tweets are, inevitably, more similar to the cluster header than the others. Our validity checks, however, show that intra-cluster correlations are very high.

**Alternative 1- Manual Coding of Content:** Manual coding has been, so far, the most commonly preferred way of conventional media analyses.<sup>40</sup> Although recent computational advances have degraded its role as a mainly employed method, researchers still rely on this labor-intensive and unsophisticated way of content analysis as a complementary method of scrutiny, especially for smaller samples, to avoid problems arising from automated coding (e.g. misspelled words not classified in standard dictionaries or misallocation of identified words based on their contextual meaning).<sup>41</sup> For larger sample sizes, however, manual coding

<sup>40</sup> Stuart Soroka, Marc André Bodet, Lori Young, and Blake Andrew, "Campaign News and Vote Intentions," *Journal of Elections, Public Opinion and Parties* 19, no. 4 (2009): 359–76; Cynthia Chew and Gunther Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 outbreak," *PloS one* 5, no. 11 (2010): e14118; Bronwyn Hemsley, Stephen Dann, Stuart Palmer, Meredith Allan, and Susan Balandin, "'We Definitely Need an Audience': Experiences of Twitter, Twitter networks and Tweet Content in Adults with Severe Communication Disabilities Who Use Augmentative and Alternative Communication (AAC)," *Disability and Rehabilitation* 37, no. 17 (2015): 1531–42.

<sup>41</sup> Chew and Eysenbach, "Pandemics in the age of Twitter"; Jessica Einspänner, Mark Dang-Anh, and Caja Thimm, "Computer-Assisted Content Analysis of Twitter Data," in *Twitter and Society*, ed. K. Weller, A. Bruns, J. Burgess, M. Mahrt, and



often falls short, either because of temporal and financial constraints or because of difficulties achieving intercoder reliability. Moreover, human coders may be prone to make subjective interpretations and misidentify certain types of latent (rather than apparent) content.<sup>42</sup> Thanks to computational tools and software, many innovative methods exist to ease incorporation of automated (quantitative) with manual (quantitative or qualitative) content analysis through identification of detectable patterns and structures of the data metrics, e.g. keyword-in-context lists (KWIC), Computer-Assisted Qualitative Data Analysis (CAQDAS), Quantitative Data Analysis (QDA).<sup>43</sup>

**Alternative 2- Fully automated coding:** Automated content coding addresses many drawbacks of manual coding and has contributed to international relations research on many topics. Automated content coding allows the computer to read data and assign each unit to pre-defined categories or categories the computer itself automatically generates. Despite substantial improvements in coding and artificial intelligence, computers may still misassign (i.e. do not understand) an important portion of the text that are fed into them. However, the fact that computers make use of very large data at virtually no cost renders statistically efficient and unbiased results.<sup>44</sup>

The structure of textual content posted on Twitter makes conducting automated content analysis quite difficult. Traditionally, automated content coding utilizes text with standard structures. For instance, news reports (including those pulled from Lexis/Nexis or Factiva-Reuters databases) exhibit a standard structure: a header and a by-line, followed by the body-text. Following conventional reporting practices, these news reports contain most of the critical information (actors, action, time and location) in the first paragraph of the body-text. “Teaching” the computer to understand sentiments and code actions is more feasible when the text is organized along such a standardized structure. In addition, these news reports are virtually free of spelling and grammatical errors.

Text from Twitter, on the other hand, is unstructured. The posts that Twitter users produce are replete with spelling and grammar mistakes, broken (quasi) sentences often due to space limitations, typing errors or simply low spelling ability of the user. To make things more complicated, online communities develop their own language, often omitting or shortening adverbs and adjectives.<sup>45</sup> Most automated methods are “restricted to simple event categories and cannot extract more complicated types of information”,<sup>46</sup> and, therefore, do not perform well in analyzing such morphed language, “obscured by the context”.<sup>47</sup> In addition, a message relayed on Twitter does not necessarily rely solely on text. Emojis, pictures and gif animations may accompany text in a tweet. Automated content coding is unable to treat such “anomalies.” Omitting such non-alphanumeric content is one option; however, such omission can lead to miscoding the content of the tweet. An eye-rolling GIF accompanying a retweet, for example, indicates that the user is most likely in disagreement with the original tweet. Manual coding after automated clustering allows us to capture such subtleties in the data.

C. Puschmann (New York, NY: P. Lang, 2014), 97–108.

<sup>42</sup> Young and Soroka, “Affective News”.

<sup>43</sup> Einspänner, et al., “Computer-Assisted Content Analysis”.

<sup>44</sup> Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston, MA: Houghton Mifflin Harcourt, 2013).

<sup>45</sup> Naomi S. Baron, *Always On: Language in an Online and Mobile World* (Oxford, UK: Oxford University Press, 2010).

<sup>46</sup> Philip A. Schrodt, “Event Data in Foreign Policy Analysis,” in *Foreign Policy Analysis: Continuity and Change in Its Second Generation*, ed. L. Neack, P. J. Haney, and J. A.K. Hey (New York, NY: Prentice Hall, 1995), 145–66.

<sup>47</sup> Grimmer and King, “General Purpose,” 5.

The lack of dictionaries and reference sets stand out as another hurdle towards using automated coding of tweets. The tweets collected for our project were in Turkish. Most automated content coding in the field of international relations (and political science) has been done for the language of English. Naturally, the “ontology” for English is much more developed than other languages, including Turkish.<sup>48</sup> While some studies have started developing such an ontology for Turkish,<sup>49</sup> these are in their infancy; building on such existing code would be too cumbersome for the purposes of this study.

Even if well-developed ontologies in Turkish existed, these could not necessarily be readily utilized. Dictionaries and reference-sets are context-dependent. They tend to -a priori- define what words a specific dimension or category include, and therefore, should be relevant to the research question at hand. Using an ontology not related to the research question at hand would be akin to having “cooking books... applied to speech in a legislature” leading to meaningless results.<sup>50</sup> Therefore, many of the reference sets either occur naturally (e.g. a war declaration for an events dataset), or otherwise rest on strong theoretical foundations (e.g. an existing party manifesto of a conservative party for coding new party positions). To develop an accurate ontology, in turn, requires a substantial amount of existing scholarly work. Many of the questions one can ask on Twitter, however, concern novel concepts expressed over unconventional media. To illustrate, one of the questions we present below asks whether security- or economics-related issues are most salient on Turkish twitter relating to Syrian refugees. The literature on public opinion of civil war refugees in host populations nascent in international relations literature; we do not have an ex-ante expectation of what categories and keywords will be of relevance for our study. This problem of coming up with relevant ontologies is further exacerbated if researchers want to associate different categories of content within a tweet. To continue with our example of Syrian refugees, a follow-up question that we ask is whether attention over a certain issue (e.g. domestic security) is associated with attitude towards refugees in a tweet. The reference set used for “issues”; however, will be quite different than the reference set used to identify whether a tweet carries a positive, negative or neutral attitude towards refugees.

### 3.2. The need for a hybrid approach: a brief technical presentation of the LCSSM and clustering

Given the unique properties of our research question, our hybrid approach stands out as one efficient way to process our data. To hand-code computer-clustered tweets for questions with immediate relevance, such as what the most prominent issues occupy the agenda on Turkish Twitter, may require substantially less time than to develop a well-working dictionary against which tweets can be individually grouped.

To calculate lexical similarity among tweets, we use a Longest Common Subsequence (LCS) similarity metric. LCS is the longest subsequence of characters from two sequences (tweets) that are common between these two sequences in the same order. This technique operates at the character level and does not infer “meaning” from text. As a result, it can operate on any alphabetic language. Figure 1 illustrates what a longest common subsequence

<sup>48</sup> Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal, “SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis,” *Twenty-eighth AAAI Conference On Artificial Intelligence* (2014): 1515–21.

<sup>49</sup> Rahim Dehkharghani, Yucel Saygin, Berrin Yanikoglu, and Kemal Oflazer, “SentiTurkNet: A Turkish Polarity Lexicon for Sentiment Analysis,” *Language Resources and Evaluation* 50, no. 3 (2016): 667–85.

<sup>50</sup> Laver, et al., “Extracting Policy Position,” 314.

is. Assume you have two sequences of characters (e.g. two tweets). If the first sequence is "thisisatest" and the second sequence is "testing123testing," then the LCS between these two sequences will be "tsitest":

thisisatest  
testing123testing  
 LCS: **tsitest**

Figure 1: Illustration of Longest Common Subsequence Metric

To illustrate technically, let's assume  $X$  and  $Y$  are two sequences and  $X_i = x_1 x_2 \dots x_i$  is the prefix of  $X$  and  $Y_j = y_1 y_2 \dots y_j$  is the prefix of  $Y$ . Then the length of the LCS can be found as follows:

$$LCS(X_i, Y_j) = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ LCS(X_{i-1}, Y_{j-1}) + 1 & \text{if } x_i = y_j \\ \max(LCS(X_{i-1}, Y_j), LCS(X_i, Y_{j-1})) & \text{if } x_i \neq y_j \end{cases}$$

The LCS problem can be solved with a dynamic programming approach in  $O(m*n)$  time and space complexity where  $m$  and  $y$  are the length of two strings. Our adaptive clustering algorithm is based on Longest Common Subsequence (LCS) as defined above. This measure, however, needs to be normalized: since different tweets will have different lengths, longer tweets are more likely to have longer common subsequences regardless of content.<sup>51</sup> To render a normalized similarity score between two tweets, we use the following formula:

$$NormalizedScore(t_i, t_j) = \frac{2 * LCS(t_i, t_j)}{Length(t_i) + Length(t_j)}$$

A naive algorithm for an LCS based tweet clustering process is given in Figure 2. This algorithm calculates the level of similarity between a single tweet and a pre-defined cluster-header tweet. If our algorithm indicates that an individual tweet is sufficiently similar to a given cluster header, then this tweet is assigned to this cluster. In other words, for each unclustered tweet (referred as tweet  $i$ ), we pass over all other following unclustered tweets (referred as tweet  $j$ ). If the normalized score that we defined above, between tweet  $i$  and tweet  $j$  is higher than the threshold, then we include indices of these tweets (which are  $i$  and  $j$ ) in the same cluster  $c$ . However, not all clusters are included into the set of clusters  $C$ ; instead only the clusters whose sizes are greater than or equal to  $k$  are included into  $C$  where  $k$  is the minimum number of tweets a cluster must have. To keep the highest level of generalizability, we set  $k=2$ , i.e. the minimum number of tweets a cluster can have is two. The first tweet ( $t_i$ ) in  $c$  becomes the representative tweet of cluster  $c$ , which is called the *cluster header*. Once we

<sup>51</sup> To iterate, our algorithm calculates similarities at the level of characters, and not at the level of words. This technique saves us the trouble of "teaching" the computer how to understand Turkish (or any language for that matter) and allows us to significant save on computing. One important shortcoming, however, is that we are unable to separate sarcastic tweets from "honest" ones in our automated coding, since both types of tweets exhibit very high levels of similarity in character content. However, since we are using manual coding for content, sarcasm does not pose a risk of validity for the two projects we present below.

include the cluster  $c$  into the set  $C$ , then we remove the tweets included in cluster  $c$  from the dataset. This algorithm ensures that the cluster sizes are greater than or equal to  $k$ .

Our technique guarantees that every tweet, which belongs to a cluster, passes the similarity threshold with the cluster header. We should also note that there is no guarantee that the similarity between any two non-cluster header tweets in a cluster is above the threshold. However, experiments show that tweets belonging to even very large clusters are similar to each other in content as well as to the representative tweet.

For our study on the Syrian refugees, we set the threshold of similarity at 0.7, leading to 82% reduction in the number of cases to be content-coded by hand. To restate, higher thresholds make it more difficult for our algorithm to decide that two tweets are alike, and therefore, cluster these tweets together. Due to the novelty of our technique, we kept this threshold relatively high. Table 1 illustrates the reduction in coding burden when the threshold is incrementally relaxed to 0.5, which increases the reduction in coding burden to 90%. Other studies that operate on more structured, clean text may decrease this threshold, and substantially reduce the time required to hand-code content of the text.

---

```

C = {};
for  $i \leftarrow 1$  to  $n$  do
  |  $t_i.isClustered \leftarrow false$ ;  $\triangleright$  Initially, all tweets are marked as unclustered;
end
for  $i \leftarrow 1$  to  $n$  do
  | if  $t_i.isClustered = false$  then
    |  $c = \{i\}$ ;
    | for  $j \leftarrow (i + 1)$  to  $n$  do
      | if  $t_j.isClustered = false$  and  $NormalizedScore(t_i, t_j) \geq threshold$ 
        | then
          |  $c \leftarrow c \cup j$ ;  $\triangleright$  Find first unclustered similar tweet
        | end
      | end
    | end
    | if  $|c| \geq k$  then  $\triangleright$  If cluster size is big enough
      |  $C \leftarrow C \cup c$ ;  $\triangleright$  Add this cluster to set of clusters
      | foreach  $index \in c$  do
        |  $t_{index}.isClustered \leftarrow true$ ;  $\triangleright$  Mark these tweets as clustered
      | end
    | end
  | end
end

```

---

Figure 2: LCS based tweet clustering algorithm

### 3.3. A few suggestions on training for aspiring IR scholars

Establishing a well-functioning relationship with computer scientists allows IR scholars a fecund way of producing interdisciplinary research. Working with computer scientists, however, does not relieve an IR scholar from the necessity of developing a basic command of coding. A computer program, like a statistics program, does not assess the quality of input; feeding garbage in will result in producing garbage out.<sup>52</sup> Furthermore, expecting a computer scientist to command the substantive content of research on discourse would be unfair at best. Therefore, possessing basic programming skills is essential for IR scholars to ensure measurement validity and consistent inference in a study. For IR scholars aspiring to obtain text manipulation techniques, two main languages of coding stand out: PERL or Python. These coding languages are mostly self-taught. A wealth of free information, tutorials, examples and communities exists online. For those who prefer to start their training in a structured way, Lubanovic<sup>53</sup> and Shaw<sup>54</sup> for Python 3, and Schwartz and Phoenix<sup>55</sup> for Perl stand out as successful introductory manuscripts. Phil Schrodt's professional website (<http://eventdata.parusanalytics.com>) contains guides and links to automated content coding software (including TABARI) and a wealth of literature on this topic. Programs mentioned in previous sections, i.e. README, *Wordscore*, *Wordfish* and many others are readily available online.

A point that needs to be underlined here is that a combination of conventional learning (i.e. learning subject by subject) and task-based hands-on learning (i.e. learning by doing) delivers the most effective results. These learning tasks can also provide (partial) answers to existing research questions. Forward-looking scholars can turn such tasks into opportunities by incorporating the outputs obtained into research papers. Contemporary computers possess sufficient computing power to conduct most textual analyses.

### 3.4. Coding, learning benefits for students and IR curricula

The benefits of learning coding far exceed enabling the collection and analysis of large amounts of data for a researcher. Such an enterprise carries many learning benefits. The process of learning coding is quite akin to learning a new language; each computer language has its own syntax wherein various constructs are logically linked to each other. The main challenge in using coding in IR arises from the need to transpose substantive theoretical questions into well-defined tasks of coding. This very challenge also constitutes the main learning benefit of coding when used in an IR curriculum. For instance, when preparing a reference dictionary, which words should be regarded as cordial or hostile in a diplomatic document? What specific army actions constitute a fortification on a border? Can we call civilians who died in a NATO bombardment as war casualties? Addressing such questions also helps the student to critically analyze what basic constructs in IR such as power, war, threat, security, alliance etc. mean, and, in addition, how these constructs logically relate to each other.

<sup>52</sup> This also points to a considerable risk associated with using off-the-shelf commercial programs. The results one obtains from such programs are very difficult to validate.

<sup>53</sup> Bill Lubanovic, *Introducing Python: Modern Computing in Simple Packages* (Sebastopol, CA: O'Reilly, 2014).

<sup>54</sup> Zed A. Shaw, *Learn Python 3 the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code* (Boston, MA: Addison-Wesley, 2017).

<sup>55</sup> Randal L. Schwartz and Tom Phoenix, *Learning Perl* (Sebastopol, CA: O'Reilly, 2001).

Incorporating coding into the IR curriculum, however, is not a straightforward task. An institutionalized response, i.e. offering a separate course in coding, surely can introduce students to the world of coding. Such a push strategy can be extremely resource-consuming in very large classes. Pooling IR students with engineering students in a conventional coding course could also substantially reduce learning benefits for IR students.

The authors of this article suggest an alternative approach: the adoption of a pull strategy for students with potential interest in coding. IR programs can easily incorporate coding across various courses across the curriculum and encourage those interested to follow through. One way to do this would be to sprinkle “enablers,” i.e. small, manageable tasks that will warm an interested student to the world of coding. Small tasks with immediate and positive feedback are often quite effective in convincing a student to delve deeper into the world of coding. Awarding extra credits for papers prepared in LaTeX is one way to entice students to employ coding in their studies. Encouraging students to start writing small bits of code for their statistical analyses (such as loops) is another such enabler. Finally, several off-the-shelf programs with easy to use graphical user interfaces exists for simple analysis. Instructors of substantive courses can easily assign tasks that require the use of these programs.

### **3.5. Personal experience of the authors**

While various institutions have started teaching courses on big-data in social sciences, neither of the social scientists authoring this study had formal education in coding. Rather, they learned methodology mostly by doing, as the projects they were responsible for required them to do so. Learning by doing tasks offers a unique opportunity; one is exposed to numerous disciplines that utilize similar methodology. Such exposure makes scholars cognizant of recent developments in these disciplines and apply these developments in international relations.

Learning by doing, however, has one important disadvantage. The programs do not necessarily work right away. In the famous Hollywood movie *Cast Away*, actor Tom Hanks plays Chuck, the protagonist who has been stranded on a remote island in the Pacific Ocean for several years. When Chuck finally decides to leave the island on a makeshift sail-raft, his first -and apparently the most formidable- challenge turns out to be crossing the large waves that break right at the coast into the open sea. Most of the obstacles in developing coding skills are akin to these breaking waves that initially prevented Chuck from leaving the island with his boat. Such obstacles occur at the very outset, especially in setting a system up and running the first simple lines of code. We came to observe that among our colleagues who give up learning code, most do so at this stage. Knowing that such problems will occur (and, alas, persist) at the outset (e.g. using a semi-colon instead of a comma in one line can prevent the running of hundreds of lines of code) can bestow mental resilience on a researcher to carry on, to cross the waves, and enjoy a smooth sail into the ocean.

## **4. Application to a Research Question: Syrian Migrants on Turkish Twitter: Which Issues Matter the Most?**

### **4.1. Motivation**

Migration, in general, and, refugee flows, in particular, have constituted a central topic of scholarly inquiry in IR over the last decade.<sup>56</sup> Turkey has been one of the countries that

<sup>56</sup> N. Katherine Hayles, “Print is Flat, Code is Deep: The importance of media-specific analysis,” *Poetics Today* 25, no. 1

received the most number of refugees, both per capita and in absolute figures, during this period. Most recent figures indicate that the number of Syrian refugees who escaped from the civil war between the Asad government and the rebels to Turkey has exceeded three million.<sup>57</sup> The social, economic and security implications of this plight inevitably placed Syrian refugees as one of the most salient issues in both in Turkish domestic politics and foreign policy.

Although public opinion has played an important role in shaping Turkey's policies towards Syrian refugees, our knowledge on what people think is scant. While a few surveys have assessed the general attitude of the Turkish public towards Syrian refugees,<sup>58</sup> we do not have systematic analysis on what specific issues the Turkish public is most concerned with. Our study aims to address this gap by utilizing Twitter data. Towards this end, we took a four-month snapshot of Turkish tweets on Syrian refugees, and assessed which issues relating to Syrian migrants turned out to be most salient on Turkish Twitter.

## 4.2. Methodology

Twitter is a popular social platform for people to share their ideas and reactions towards daily events. Analyzing tweets sent for a specific topic is important since these tweets give important insight to online public opinion.<sup>59</sup> However, analyzing large volume of tweets in an unbiased way is a challenging task. Conventional hand-coding is simply unfeasible for very large data. Automated sentiment analysis and machine coding of text is another option.<sup>60</sup> Language used on Twitter, however, is “dirty” (i.e., unstructured), hence very difficult to form a relevant “ontology” from which a computer can learn. Instead we use the CLS similarity metric technique detailed in the previous section.

In this study, our novel clustering technique allowed 52845 of the 60146 tweets to be grouped into 3553 clusters, while 7301 of these tweets turned out as stand-alone tweets. These 10874 (3553 cluster headers + 7321 remaining stand-alone tweets) individual entries were, then, coded for content by the authors. Further statistical analyses were conducted on these groups to gauge the relationship between various aspects of the content of the tweets. Clustering led to a decrease of the number of entries that needed to be hand-coded dataset by a factor of 5.5.

## 4.3. Data Collection

We started to collect real time Turkish tweets by using Twitter Stream API, which contain the keyword of *mülteci* [refugee], including the spelling with English letters, i.e. *multeci* as well as the word *Suriye* and its inflections in Turkish (such as *Suriyeli*, *Suriyeliler* etc.). The collected tweets were, then, filtered for language; this filter eliminated tweets that were not

(2004): 67–90.

<sup>57</sup> “Suriyeli sığınmacılara yapılan yardımlar,” AFAD, <https://www.afad.gov.tr/tr/2373/Suriyeli-Siginmacilara-Yapilan-Yardimlar> (2017).

<sup>58</sup> See for example EDAM, “Reaction Mounting against Syrian Refugees in Turkey,” Public Opinion Surveys of Turkish Foreign Policy no:1 (2014); The German Marshall Fund of the United States, *Turkish Perceptions Survey* (2015); M. Murat Erdoğan, *Türkiye'deki Suriyeliler: Toplumsal kabul ve uyum* (İstanbul: İstanbul Bilgi Üniversitesi Yayınları, 2015).

<sup>59</sup> Gökçe, et al., “Twitter and Politics”; Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe, “Predicting Elections with Twitter: What 140 characters reveal about political sentiment,” *JCWSM* 10, no. 1 (2010): 178–85; Presley Ifukor, “Elections” or “Selections”? Blogging and Twittering the Nigerian 2007 General Elections,” *Bulletin of Science, Technology & Society* 30, no. 6 (2010): 398–414.

<sup>60</sup> Philip A. Schrodt, “Precedents, Progress, and Prospects In Political Event Data,” *International Interactions* 38, no. 4 (2012): 546–69.

in Turkish language. As a result, our analysis comprises of 63312 tweets, which were posted between 2 May 2016 and 25 August 2016. An important limitation at this step was Twitter's limited access for legal retrieval of related tweets via Twitter Stream API upon a word query. The true number of tweets that contained the word *mülteci and Suriye* are much higher than 63312. Nonetheless, we have no apriori reason to believe that Twitter censors results from query in a non-random way.<sup>61</sup> Therefore, our sample is sufficiently large for the inferences made below.

#### 4.4. Clustering and coding of tweets

Standard document clustering methods in data mining do not work well on Twitter data due to several challenges explicated in the previous section. To address these challenges, we utilized an algorithm based on the Least Common Subsequence similarity metric, which utilizes an adaptive clustering approach for Twitter which is based on lexical similarity.<sup>62</sup> Based on our algorithm, 52845 of the 63312 tweets clustered successfully under 3553 clusters, while 7301 unclustered individual tweets remained. We, then, hand coded these 10854 (3553+7301) entries for various content including the issues mentioned in the tweet, attitudes towards refugees, whether any political party is mentioned and attitudes towards this political party, mention of any terrorist organizations, among others. Four separate coders were utilized in this study. 500 randomly chosen tweets were jointly coded to measure intercoder reliability. The lowest pairwise correlation among any pair of coders was 0.88.

#### 4.5. Findings

Our survey of related literature, policy and news reports led us to establish ten issue topics which might identify a given cluster header: (1) security within the borders of Turkey (TR Security); (2) security outside of Turkey's borders (Security Abroad); (3) immigrants' safety; (4) economy; (5) social aid; (6) identity/ethnicity; (7) granting of Turkish citizenship (TR Citizenship); (8) demographical change; (9) general (for headers without specific topic); and, (10) other (for headers not having any connection with Syrian-refugee issues). These categories are not mutually exclusive; a tweet can refer to multiple issues.

Table 1- Salience of Issues Concerning Syrian Refugees on Turkish Twitter

	Frequency	Percentage	TOTAL
<b>TR Security</b>	22,102	34.91	63,312
<b>Security Abroad</b>	2,931	4.63	63,312
<b>Immigrants' Safety</b>	13,327	21.05	63,312
<b>Economy</b>	8,775	13.86	63,312
<b>Social Aid</b>	4,614	7.29	63,312
<b>Identity/Ethnicity</b>	5,116	8.08	63,312
<b>TR Citizenship</b>	10,104	15.96	63,312
<b>Demography</b>	758	1.2	63,312
<b>General</b>	9,703	15.33	63,312
Other	4,947	7.81	63,312

<sup>61</sup> Twitter does not publicly disclose the sampling mechanism; however, the company officially declares that "the Sample endpoint provides a statistically relevant sample of 1% of the full firehose." See <https://twittercommunity.com/t/is-the-sample-streaming-api-truly-random/14942/3>

<sup>62</sup> İnanç Arın, Mert Kemal Erpam and Yücel Saygın, "I-TWEC: Interactive Clustering Tool for Twitter," *Expert Systems with Applications* 96 (2018): 1-13.



Table 1 lists the frequencies and relative percentages of the issues mentioned in the 63,312 tweets analyzed for this study. Our findings indicate concern for domestic security in Turkey seems to trump over others on Twitter. Many of the tweets grouped under this category highlighted the government's inability to protect its citizens. Some also suggested a global conspiracy against Turkey and portrayed the inflow of Syrian refugees as a part of a bigger plan to "bring Turkey to its knees." Given that Turkey had been suffering several terrorist attacks during 2015 and 2016, many of which were linked to various groups involved in the Syrian civil war, this result is not surprising. Nonetheless, establishing that the Syrian refugees issue has been strongly linked to domestic security concerns constitutes a solid basis for further studies.

The safety of immigrants turns out to be the second most important topic in our sample of tweets; more than one in every five tweets mention the problems immigrants face, in Turkey or elsewhere. Further analyses of our data also indicate that almost all of the tweets that mention immigrants' safety take on a positive attitude towards the immigrants themselves. Hence, we can safely conclude that a significant part of Turkish online community care about the physical and psychological integrity of immigrants. Many of these tweets focused on the allegations of sexual abuse of refugee children, both in the cities and the camps. A significant number also drew attention to the government's prohibition of third party probe into these allegations. A much smaller, but still important, group drew attention to the plight of the refugees escaping from the violence of the Assad regime.

The granting of citizenship to Syrian refugees occupied the third most salient place on the agenda. Turkish twitter community was divided on this topic. One camp highlighted the adverse security implications of granting citizenship to Syrians and argued that Syrian refugees as Turkish citizens would constitute a fifth column for Turkish society. The other camp motivated the case for granting citizenship to Syrian refugees based on religious kinship. A general theme that repeated itself in this camp was Syrian refugees deserved citizenship as much as other groups to whom Turkey has given citizenship, such as refugees from the Balkan, Armenians, Jews and seculars. Interestingly, citizenship and pecuniary issues were not paired as frequently.

The impact Syrian refugees have on Turkish economy turned out to be the fourth most salient specific issue area. Tweets falling under this category were predominantly organized under three main arguments. The first argument highlighted the overall cost on Turkish economy. A second group of tweets directed their criticism towards the redistributive consequences of this impact and argued that those in favor of the inflow of refugees should be the ones shouldering the economic burden these refugees bring. The third line of argument directed criticism against the critics themselves. These users noted that other countries, such as Germany and Jordan, were also suffering as much economically, but did not complain as much from the inflow of Syrian refugees on economic grounds.

Among the tweets that identified at least one specific policy issue, the tweets underlying the ethnic- or identity related dimensions constitute the fifth biggest category. Many of these tweets grouped under this category are also associated with another policy issue. Still, the prevalence of ethnic-/identity-based evaluations of policy options warrant ethnicity/identity to be treated as a separate group. One significant subgroup of tweets under this category criticized the preferential treatment of Syrian over Iraqi Turkmen refugees. Another interesting strand of tweets saw the inflow of Syrian refugees as a grand plan of Armenia and

the United States to subjugate Turkey. The prevalence of such ethnic concerns also suggests that competing identities within Turkey may be shaping an individual's outlook towards the issue of Syrian refugees. Reflecting on some of the most agitating reports on popular media, we decided to code concerns regarding demographic change the settling of Syrian refugees may cause as an issue separate from identity/ethnicity. While not emerging as one of the most important topic, a notable 1.2% of the tweets expressed concern about potential ethnic tensions. Most of these tweets asserted that many Sunni Syrian refugees were knowingly being settled into villages that are predominantly Alevi.

A related topic to economy, social aid turned out to be the sixth most important issue area. A brief survey of conventional media highlight the strain Syrian refugees put on education, hospitals and other public services. Our findings showed that Turkish twitters give little attention to the relationship between Syrian refugees and public services in Turkey. As a matter of fact, what very little is talked about healthcare and education take on a positive tone. The most prominent debate on social aid, however, revolved around the government's promise to transfer the Authority for Public Housing (TOKI) apartment units to Syrian refugees free of charge. Most tweets on this topic were of critical tone, often noting the difficulties Turkish military veterans have in obtaining housing from the state.

## 5. Conclusion, Tips and Warnings

In her treatise on emerging forms of text as alternatives to print, Hayles states that “nature of the medium in which [texts] are instantiated matters.”<sup>63</sup> Indeed, the nature of these media also matters for the research design scholars adopt in analyzing the text. This study presents a novel technique developed by the I-POST team at Sabancı University. The technique is easy to use, scalable and applicable to all alphabetical languages. Our brief presentation of findings from two studies establish the validity of our technique.

We argue that CLS similarity metric clustering strikes a good balance between the high-levels of upfront cost in developing a strong infrastructure for fully automated coding and the burden and unreliability of manual coding. While it reduces the burden of manual coding (up to 90%), it also allows the coders to decide this technique is especially useful in taking public thermometers in fast-moving environments where the language itself also evolves very quickly, such as Twitter.

Twitter offers large swathes of opinion data on a wide array of topics. Computer-assisted data analysis allows researchers to unlock this vast data source. However, a number of caveats should be kept in mind regarding (i) using Twitter data and (ii) employing computer-assisted data analysis techniques. This paper does not talk about techniques relating to scraping data off the internet. Another caveat relates to the homogeneity of the clusters. Our clustering mechanism matches each tweet that passes the preset threshold with the cluster header tweet. While our code guarantees each cluster-header – regular tweet pair to exceed the preset threshold within a cluster, it does not guarantee such a level of similarity for each tweet-pair within a cluster. That said, the intra-correlation among individual tweets within each cluster turned out to be quite high in our project. Such an intra-correlation test should be conducted for each project. Note that, this study is also important to give a snapshot of relative prevalence of topics in the issue space on Syrian refugees in Turkey. As such the unit

<sup>63</sup> Hayles, “Print is Flat”.

of analysis is “tweet,” and not user. While some studies have shown that voting behavior can be estimated from Twitter data reasonably well,<sup>64</sup> estimating public opinion representative of the population (à la surveys) from Twitter data falls outside of the scope of our analysis.

The effect bots (i.e. tiny scripts that post content, imitating a real person) and trolls (real people posting with the explicit aim of detracting or reframing online debates using various techniques of rhetoric) may have on the conclusions of studies using Twitter data also merit further discussion. Bots and trolls tend to neutralize, reframe and/or spin the discourse the debate on existing issues; their presence would only augment the count for issues for which the government is under criticism. Recent studies indicate that bots comprise 9-15% of Twitter users who post in English.<sup>65</sup> Assuming a similar proportion (if not less) holds for Twitter bot accounts tweeting in Turkish, the presence of such accounts is unlikely to change our substantive findings on issue prevalence presented in this study; domestic security seems to trump over all other concerns on Turkish Twitter.

Trolls and bots may be of a more serious concern for studies that correlate various aspects of tweet content, such as support for granting citizenship to Syrian refugees and attitudes towards the incumbent party. If bots and trolls are more readily employed by those favoring the incumbent party, then our sample would overrepresent the proportion of tweets that (i) are in favor of government policies, (ii) express a positive attitude towards the refugees, and/or (iii) downplay the difficulties refugee inflows have brought about in Turkish society. Therefore, the number of “pro-government” tweets would increase for topics that normally would draw criticism to the government. Then, for those who hypothesize that certain issues are mentioned more by government critics, such biases are conservative biases, that is, the presence of trolls and bots make it more difficult to derive statistically significant correlations. Our further analyses, not presented in this paper, find that tweets criticizing the issues brought about by refugee inflows tend to be critical of the government. Future studies could look at in which political issues bots and trolls are more active or identify bot and troll accounts and re-conduct analyses by removing the posts by these accounts.

A final caveat that needs to be kept in mind relates to temporal shocks to the data collection. Our four-month time window for data collection witnessed two such shocks. The first shock was the blowing up of two Syrian refugees in a dwelling in Hatay province while they were preparing explosives. This occurred within a few days of a public announcement by President Erdogan, which suggested Turkey should grant Turkish citizenship for “qualified” Syrians. The second shock was the failed coup attempt of July 15, 2016, which crowded most of the tweets on Syrian refugees out from the political agenda on Turkish twitter. We argue that such temporal shocks do not necessarily threaten the validity of our findings. One could initially expect the bomb-incident to be the last straw to pour out latent resentment against the economic implications of Syrian refugee inflows to Turkey. Surprisingly, though, as President Erdogan’s “qualified labor” statement was echoed in most of the tweets after the blow-up of the bomb, the citizenship dimension—and the effect granting citizenship would have on future domestic terrorism—was recalled significantly more than the economic/unemployment dimension.

<sup>64</sup> Pablo Barberá, “Less is More? How Demographic Sample Weights Can Improve Public Opinion Estimates Based On Twitter Data” (working paper, NYU, 2016); Tumasjan, et al., “Predicting Elections”.

<sup>65</sup> Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini, “Online Human–Bot Interactions: Detection, Estimation, and Characterization,” arXiv:1703.03107 (2017).

We further argue that such unexpected events allow researchers to probe further into evolution of political sentiments on Twitter. An interesting question, for instance, is whether the failed coup attempt created a rally-around-the-flag effect and allowed the government some breathing room against criticisms for its Syrian refugee policy on the internet. Scholars can also plan for scheduled events such as elections, international summits and start collecting data sufficiently before such events to gather statistically meaningful data.

A final point to note about social media messages are rich media that offer content more than text. For example, further data can be obtained from images used and/or linked to in Twitter messages by employing simple crawling and query code. One possibility could be to gather immediate Google search results as text data for each image queried. CLS metric, then, can allow clustering of these images and allow for systematic analyses based on images. Revisiting image research studies of foreign policy analysis<sup>66</sup> from a large-N analysis could render interesting results.

## Bibliography

- Adler, E. Scott, and John D. Wilkerson. *Congress and the Politics of Problem Solving*. Cambridge, UK: Cambridge University Press, 2013.
- Arın, İnanç. *Impact Assessment & Prediction of Tweets and Topics*. Ph.D. thesis, Sabancı University, 2017.
- Arın, İnanç, Mert Kemal Erpam, and Yücel Saygın. "I-TWEC: Interactive Clustering Tool for Twitter." *Expert Systems with Applications* 96 (2018): 1–13.
- Barberá, Pablo. *Less is More? How Demographic Sample Weights Can Improve Public Opinion Estimates Based on Twitter Data*. Working paper, NYU, 2016.
- Baron, Naomi S. *Always On: Language in an Online and Mobile World*. Oxford, UK: Oxford University Press, 2010.
- Bollen, Johan, Huina Mao, and Alberto Pepe. "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena." *ICWSM* 11 (2011): 450–53.
- Bond, Doug, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development." *Journal of Peace Research* 40, no. 6 (2003): 733–45.
- Braithwaite, Alex, and Douglas Lemke. "Unpacking Escalation." *Conflict Management and Peace Science* 28, no. 2 (2011): 111–23.
- Cambria, Erik, Daniel Olsher, and Dheeraj Rajagopal. "SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis." *Twenty-eighth AAAI Conference on Artificial Intelligence* (2014): 1515–21.
- Chew, Cynthia, and Gunther Eysenbach. "Pandemics in the Age of Twitter: Content Analysis of Tweets During The 2009 H1N1 Outbreak." *PLoS one* 5, no. 11 (2010): e14118.
- Cottam, Martha L. *Foreign Policy Decision Making: The Influence of Cognition*. Boulder, CO: Westview Press, 1986.
- Cull, Nicholas J. "The Long Road to Public Diplomacy 2.0: The Internet in US Public Diplomacy." *International Studies Review* 15, no. 1 (2013): 123–39.
- de Jong, Ms Wilma, Martin Shaw, and Neil Stammers. *Global Activism, Global Media*. London, UK: Pluto Press, 2005.
- Dehkharghani, Rahim, Yücel Saygın, Berrin Yanikoglu, and Kemal Oflazer. "SentiTurkNet: a Turkish Polarity

<sup>66</sup> See for example Martha L. Cottam, *Foreign Policy Decision Making: The Influence of Cognition* (Boulder, CO: Westview Press, 1986); Richard K. Herrmann and Michael P. Fischerkeller, "Beyond the Enemy Image and Spiral Model: Cognitive-Strategic Research after the Cold War," *International Organization* 49, no. 3 (1995): 415–50.

- Lexicon for Sentiment Analysis.” *Language Resources and Evaluation* 50, no. 3 (2016): 667–85.
- Demirhan, Kamil. “Social Media Effects on the Gezi Park Movement in Turkey: Politics under Hashtags.” In *Social Media in Politics: Case Studies on the Political Power of Social Media*, edited by Bogdan Pătruț and Monica Pătruț, 281–314. Basel: Springer International Publishing, 2014.
- Dugan, Laura, and Erica Chenoweth. “Government Actions in Terror Environments (GATE): A Methodology that Reveals How Governments Behave toward Terrorists and Their Constituencies.” In *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian, 465–86. New York, NY: Springer, 2013.
- Einspanner, Jessica, Mark Dang-Anh, and Caja Thimm. “Computer-Assisted Content Analysis of Twitter Data.” In *Twitter and Society*, edited by K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann, 97–108. New York, NY: P. Lang, 2014.
- Erdoğan, M. Murat. *Türkiye'deki Suriyeliler: toplumsal kabul ve uyum*. İstanbul Bilgi Üniversitesi Yayınları, 2015.
- Erdoğan, İrem Eren, and Mesut Cicek. “The Impact of Social Media Marketing on Brand Loyalty.” *Procedia-Social and Behavioral Sciences* 58 (2012): 1353–60.
- Gerner, Deborah J., Philip A. Schrod, Omür Yılmaz, and Rajaa Abu-Jabr. “Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions.” Paper presented at the Annual Meeting of the International Studies Association, New Orleans, 2002.
- Ghosn, Faten, Glenn Palmer, and Stuart A. Bremer. “The MID3 Data Set, 1993–2001: Procedures, Coding Rules, and Description.” *Conflict Management and Peace Science* 21, no. 2 (2004): 133–54.
- Gökçe, Osman Z., Emre Hatipoğlu, Gökhan Göktürk, Brooke Luetgert, and Yücel Saygın. “Twitter and Politics: Identifying Turkish Opinion Leaders in New Social Media.” *Turkish Studies* 15, no. 4 (2014): 671–88.
- Goldstein, Joshua S. “A Conflict-Cooperation Scale for WEIS Events Data.” *Journal of Conflict Resolution* 36, no. 2 (1992): 369–85.
- Grimmer, Justin, and Gary King. “General Purpose Computer-Assisted Clustering and Conceptualization.” *Proceedings of the National Academy of Sciences* 108, no. 7 (2011): 2643–50.
- Grimmer, Justin, and Brandon M. Stewart. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21, no. 3 (2013): 267–297.
- Hatipoğlu, Emre, Osman Z. Gökçe, Berkay Dinçer, and Yücel Saygın. “Sosyal medya ve Türk dış politikası: Kobani Tweetleri üzerinden Türk dış politikası algısı.” *International Relations/Uluslararası İlişkiler* 13, no. 52 (2016): 175–97.
- Hayles, N. Katherine. “Print is Flat, Code is Deep: The Importance of Media-Specific Analysis.” *Poetics Today* 25, no. 1 (2004): 67–90.
- Hemsley, Bronwyn, Stephen Dann, Stuart Palmer, Meredith Allan, and Susan Balandin. “‘We Definitely Need an Audience’: Experiences of Twitter, Twitter Networks and Tweet Content in Adults with Severe Communication Disabilities Who Use Augmentative and Alternative Communication (AAC).” *Disability and Rehabilitation* 37, no. 17 (2015): 1531–42.
- Herrmann, Richard K., and Michael P. Fischerkeller. “Beyond the Enemy Image and Spiral Model: Cognitive-Strategic Research after the Cold War.” *International Organization* 49, no. 3 (1995): 415–50.
- Hjorth, Frederik, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. “Computers, Coders, and Voters: Comparing Automated Methods for Estimating Party Positions.” *Research & Politics* 2, no. 2 (2015): 1–9.
- Hopkins, Daniel J., and Gary King. “Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability.” *Public Opinion Quarterly* 74, no. 2 (2010): 201–22.
- Ifukor, Presley. “‘Elections’ or ‘Selections’? Blogging and Twittering the Nigerian 2007 General Elections.” *Bulletin of Science, Technology & Society* 30, no. 6 (2010): 398–414.
- Kalathil, Shanthy, and Taylor C. Boas. *Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule*. Washington, DC: Carnegie Endowment, 2010.
- Kenwick, Michael R., Matthew Lane, Benjamin Ostick, and Glenn Palmer. “Codebook for the Militarized Interstate

- Incident Data, version 4.0." Unpublished manuscript, 2013.
- King, Gary, Patrick Lam, and Margaret E. Roberts. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61, no. 4 (2017): 971–88.
- Kissinger, Henry. *Diplomacy*. New York, NY: Simon and Schuster, 1994.
- Klemmensen, Robert, Sara Binzer Hobolt, and Martin Ejnar Hansen. "Estimating Policy Positions Using Political Texts: An Evaluation of the Wordscores Approach." *Electoral Studies* 26, no. 4 (2007): 746–55.
- König, Thomas, Brooke Luetgert, and Tanja Dannwolf. "Quantifying European legislative Research: Using CELEX and PreLex in EU legislative studies." *European Union Politics* 7, no. 4 (2006): 553–74.
- LaFree, Gary, Laura Dugan, and Erin Miller. *Putting Terrorism in Context: Lessons from the Global Terrorism Database*. London, UK: Routledge, 2014.
- Laver, Michael, and Kenneth Benoit. "Locating TDs in Policy Spaces: The Computational Text Analysis of Dáil Speeches." *Irish Political Studies* 17, no. 1 (2002): 59–73.
- Laver, Michael, Kenneth Benoit, and John Garry. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97, no. 2 (2003): 311–31.
- Lubanovic, Bill. *Introducing Python: Modern Computing in Simple Packages*. Sebastopol, CA: O'Reilly, 2014.
- Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt, 2013.
- McClelland, Charles A., Rodney G. Tomlinson, Ronald G. Sherwin, Gary A. Hill, and Herbert L. Calhoun. *The Management and Analysis of International Event Data: A Computerized System for Monitoring and Projecting Event Flows*. University of Southern California Los Angeles School of International Relations, 1971.
- Melissen, Jan, and Emillie V. de Keulenaar. "Critical Digital Diplomacy as a Global Challenge: The South Korean Experience." *Global Policy* 8, no. 3 (2017): 294–302.
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. "Fightin'words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16, no. 4 (2008): 372–403.
- Organski, A.F.K., and Jacek Kugler. *The War Ledger*. London, UK: University of Chicago Press, 1981.
- Palmer, Glenn, Vito d'Orazio, Michael Kenwick, and Matthew Lane. "The MID4 dataset, 2002–2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science* 32, no. 2 (2015): 222–42.
- Pötzsch, Holger. "The Emergence of iWar: Changing Practices and Perceptions Of Military Engagement in a Digital Era." *New Media & Society* 17, no. 1 (2015): 78–95.
- Proksch, Sven-Oliver, and Jonathan B. Slapin. "WORDFISH: Scaling Software for Estimating Political Positions From Texts." *Version 1* (2008): 323–44.
- Purdy, Christopher H. "Using the Internet and Social Media to Promote Condom Use in Turkey." *Reproductive Health Matters* 19, no. 37 (2011): 157–65.
- Schrodt, Philip A. "Event Data in Foreign Policy Analysis." In *Foreign Policy Analysis: Continuity and Change in Its Second Generation*, edited by L. Neack, P. J. Haney, and J. A.K. Hey, 145–66. New York, NY: Prentice Hall, 1995.
- \_\_\_\_\_. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38, no. 4 (2012): 546–69.
- Schrodt, Philip A., and Deborah J. Gerner. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982–92." *American Journal of Political Science* (1994): 825–54.
- Schrodt, Philip A., Glenn Palmer, and Mehmet Emre Hatipoğlu. "Automated Detection of Reports of Militarized Interstate Disputes Using the SVM Document Classification Algorithm." Paper presented at the Annual Meeting of the American Political Science Association, Boston, 2008.
- Schrodt, Philip A., Shannon G. Davis, and Judith L. Weddle. "Political Science: KEDS- a Program for the Machine Coding of Event Data." *Social Science Computer Review* 12, no. 4 (1994): 561–87.
- Schwartz, Randal L., and Tom Phoenix. *Learning Perl*. Sebastopol, CA: O'Reilly, 2001.
- Seib, Philip. *Real-Time Diplomacy: Politics and Power in the Social Media Era*. New York, NY: Palgrave Macmillan

- US, 2012.
- Shaw, Zed A. *Learn Python 3 the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code*. Boston, MA: Addison-Wesley, 2017.
- Shellman, Stephen M. "Coding Disaggregated Intrastate Conflict: Machine Processing The Behavior Of Substate Actors over Time and Space." *Political Analysis* 16, no. 4 (2008): 464–77.
- Small, Melvin, and Joel David Singer. *Resort to Arms: International and Civil Wars, 1816-1980*. London, UK: Sage Publications, Inc, 1982.
- Soroka, Stuart, Marc André Bodet, Lori Young, and Blake Andrew. "Campaign News and Vote Intentions." *Journal of Elections, Public Opinion and Parties* 19, no. 4 (2009): 359–76.
- Tüfekçi, Zeynep. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. New Haven, CT: Yale University Press, 2017.
- Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10, no. 1 (2010): 178–85.
- Ünver, H. Akın, and Hassan Alassaad. "How Turks Mobilized against the Coup: The Power of the Mosque and the Hashtag." *Foreign Affairs*, September 14, 2016.
- Varol, Onur, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. "Online Human-Bot Interactions: Detection, Estimation, and Characterization." *arXiv:1703.03107* (2017).
- Yavuz Gökem, Şenay. "The Weakest Link or the Magic Stick?: Turkish Activists' Perceptions on the Scope and Strength of Digital Activism." *Turkish Studies* 18, no. 1 (2017): 102–24.
- Young, Lori, and Stuart Soroka. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29, no. 2 (2012): 205–31.