Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis

Alan Mishler(✉), Erin Smith Crabb, Susannah Paletz, Brook Hefright, and Ewa Golonka

University of Maryland, College Park, MD, USA
{amishler, ecrabb, paletz, hefright, egolonka}@umd.edu

Abstract. Structural topic modeling (STM) is a recently introduced technique to model how the content of a collection of documents changes as a function of variables such as author identity or time of writing. We present two proof-of-concept applications of STM using Russian social media data. In our first study, we model how topics change over time, showing that STM can be used to detect significant events such as the downing of Malaysia Air Flight 17. In our second study, we model how topical content varies across a set of authors, showing that STM can be used to cluster Twitter users who are sympathetic to Ukraine versus Russia as well as to cluster accounts that are suspected to belong to the same individual (so-called "sockpuppets"). Structural topic modeling shows promise as a tool for analyzing social media data, a domain that has been largely ignored in the topic modeling literature.

Keywords: Structural topic modeling · Event detection · Authorship attribution · Public opinion measurement · Social media

1 Introduction

Topic modeling is a statistical and computational technique for discerning information about the contents of a large corpus of documents without reading or annotating the original texts [1]. A topic model uncovers patterns of word co-occurrence across the corpus, yielding a set of word clusters, together with associated probabilities of occurrence, which constitute the 'topics'.

Topic modeling is of interest to any user who wishes to gain insight into a collection of documents, including researchers who want to understand what is being discussed online. While the use of topic modeling is well attested with texts such as novels or news stories [2, 3], relatively little work has been done in the realm of social media. Modeling in this domain is often more challenging given character limits, which cause users to condense their messages in a variety of ways, such as by using URL shortening services when posting links [4].

The standard topic modeling technique, Latent Dirichlet Allocation (LDA), may have limited utility in the realm of social media. LDA makes a statistical assumption that all texts in the modeled corpus are generated by the same

C. Stephanidis (Ed.): HCII 2015 Posters, Part I, CCIS 528, pp. 639–644, 2015.
DOI: 10.1007/978-3-319-21380-4_108

underlying process [5]. Thus, it is not ideally suited to examining differences in topical content that are affected by external variables such as author identity or time of writing.

Structural topic modeling (STM) is a recently introduced variant of LDA that is designed to address precisely this limitation [6]. STM can represent the effect of external variables on both topical content and topical prevalence. Topical content refers to the probabilities associated with words in each topic, while topical prevalence refers to the proportions of different topics that occur within documents. The external variables can consist of any metadata that distinguishes one text from another, including variables relating to author identity (gender, age, political affiliation, etc.), textual genre (for example, news stories versus academic articles), and time of production.

Since STM is a relatively recent innovation, its full utility has not yet been well demonstrated. We investigated two particular applications of STM related to Russian and Ukrainian social media data. In the first study, we show that STM can be used to detect major real world events such as the downing of Malaysia Air Flight 17 (MH 17) in Ukraine. In the second study, we demonstrate that STM can be used to group and distinguish Twitter users with different political sympathies, namely, those sympathetic to Russia versus Ukraine.

2 Study 1: Event Detection

2.1 Methods

We downloaded 50,000 posts from VKontakte, a Russian-owned social media site similar to Facebook. The posts were gathered using TweetTracker, a web-based portal for collecting tweets and other social media data [7, 8]. As a filter for selection, we used the "VK:Ukraine" task, a publicly available search filter designed to gather data pertaining to Ukraine. Specifically, we harvested 25,000 posts from July 16, 2014, the day before the MH 17 crash, and 25,000 posts from July 18, 2014, the day after the crash. We gave the posts a label of "t1" (Time 1) and "t2" (Time 2), respectively. In order to ensure that the model did not simply identify clusters of words based on the source language, we collected posts in a single language, Russian, as determined by the Google Translate language code included in the TweetTracker metadata.

We constructed a 30-topic structural topic model of the data using the stm package in R [9]. Prior to modeling, strings beginning with "#", "@", and "http" were removed, as were non-UTF-8-encodable characters and function words such as "and" and "the." Words were stemmed using the *textProcessor* function in the stm package. Since the goal was to detect a change in expressed content from Time 1 to Time 2, we conditioned both topical content (the probabilities associated with words in each topic) and topical prevalence (the proportions of different topics that occur within documents) on the two-level time covariate ("t1" versus "t2").

2.2 Findings

Topical prevalence differed substantially from Time 1 to Time 2 for nearly all of the 30 topics modeled (Fig. 1). In particular, two topics, identified arbitrarily as Topic 2 and Topic 13, were significantly more common in the corpus at Time 2 than at Time 1.

An examination of prominent words within each topic (Table 1) makes it clear that these topics relate to the MH 17 crash and its purported cause – namely, the Buk missile system that is believed to have been used to shoot down the plane [10].

As predicted, the model easily detected changes in content that resulted from the MH 17 crash. A researcher with minimal training in interpreting the results of topic models and no prior knowledge of the MH 17 event would be able to readily infer from these results that a catastrophe involving a plane and a missile system occurred between

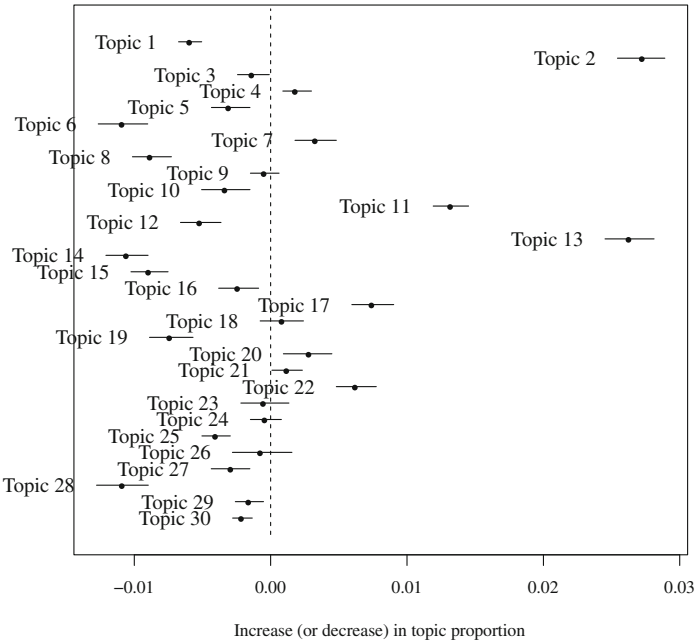

Fig. 1. Difference in topic prevalence from Time 1 to Time 2 for each of the 30 topics modeled in study 2A. A positive value indicates that the topic is more prevalent at Time 2; a negative value indicates that the topic is more prevalent at Time 1.

Table 1. Study 1: prominent words in Topic 2 and Topic 13. These two topics are significantly more prevalent in the corpus at Time 2 than at Time 1.

Topic	Prominent words
Topic 2	*самолет* 'airplane', *боинг* 'Boeing', *бук* 'Buk (missile system)'
Topic 13	*Расследован* 'investigated', *катастроф(а)* 'catastrophe', *малайзийск(ий)* 'Malaysian', *борт* 'board'

July 16 and July 18. The researcher could then examine a subset of the actual VK posts or other documents to determine the precise nature of the event.

3 Study 2: Clustering Twitter Users

3.1 Methods

The dataset for this study consisted of approximately 4,000 Russian language tweets collected via the "Crimea" TweetTracker query. The "Crimea" query returns tweets that include a key word or hashtag referencing a set of pre-defined key words pertaining to Crimea or Ukraine, come from a particular set of users related to Crimea, or are geotagged as originating from the area defined by a geographic coordinate bounding box in the Crimea region. The tweets were timestamped between June 25 and August 25, 2014, which was after the (previously Ukrainian) region of Crimea joined the Russian federation on March 18, 2014 [11]. This was a period of continued tension between Russia and Ukraine, and we expected that a high proportion of tweets would be about this crisis. (Note that this is a subset of the dataset used for the authorship attribution study reported in [12] in this same volume).

The tweets came from four users, labeled arbitrarily as S, K, L, and T. Users K and L were selected in part because the authorship attribution analysis reported in [12] found these two users to be highly similar. That analysis entailed using character bigrams to quantitatively assess the uniqueness of users' sets of tweets. If STM also finds evidence of these users' similarity, then the results of these two methods will be mutually validating. In other words, the conclusion that these two users are highly similar will be strengthened, and the confidence in each method will also be strengthened.

An additional goal with this model was to determine whether STM can be used to group and distinguish these users from one another. We added users S and T toward this end. Users S, K, and L were selected because two Russian linguists on the research team judged them to be generally sympathetic to Ukraine, while we selected user T because the linguists judged that user to be generally sympathetic to Russia and hostile to Ukraine. The linguists based their judgments on the content and tone of a sample of several hundred tweets from each user. They did not know the results of the analysis prior to reading the tweets and were instructed to simply characterize in their own words any similarities and differences among the users.

We preprocessed the data in the same manner as in Study 1 and modeled it again using the stm package in R [9]. Due to the small size of the corpus and the fact that tweets are limited to 140 characters, we modeled only five topics (as compared to 30 in Study 1). We conditioned topical prevalence on the four-level factor author identity, with each author constituting a factor level.

3.2 Findings

Plots of topic prevalence by author show that tweets from users S, L, and K are topically similar, while the content of user T's tweets is markedly different.

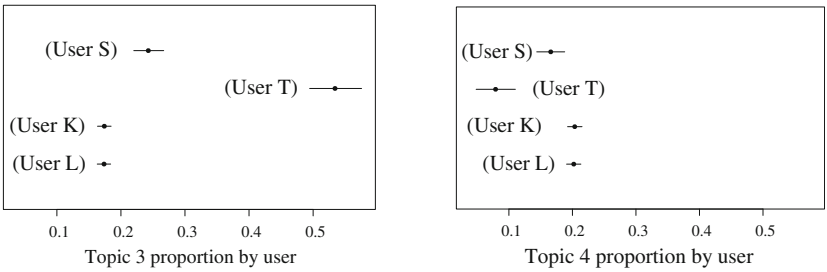

Fig. 2. Study 2: Proportion of words from Topics 3 and 4 for each user, across all of that user's tweets.

Table 2. Study 2: prominent words in Topics 3 and 4

Topic	Prominent words
Topic 3	*США* 'USA', *границ(ы)* 'borders', *Крым* 'Crimea', *Росс(ия)* 'Russia', *Путин* 'Putin'
Topic 4	*Донецк* 'Donetsk', *Луганск* 'Luhansk', *обстрел* '[arms] fire', *град* 'a multiple rocket launcher system', *погиб(ли)* 'died', *силовик* 'political actor associated with Russian special services'

In particular, user T relies substantially on Topic 3, while the other four topics are less prevalent in user T's tweets than in the other users' tweets (Fig. 2; again, the indices used to identify topics are arbitrary). Topics 1, 2, and 5 are not plotted, but they have similar distributions to that of Topic 4.

These results confirm and strengthen the conclusion from the sockpuppet detection analysis reported in [12] that users K and L are highly similar. Additionally, they reveal the expected two clusters of users: the three users sympathetic to Ukraine versus the one user sympathetic to Russia. Prominent words from Topics 3 and 4 give some indication of how the content of these users' tweets differs (Table 2). Notably, Topic 4 contains references to weaponry, death, and Russian special services, while Topic 3 appears to be primarily related to geographic locations and geopolitical actors.

A researcher or analyst with no prior knowledge of the content of these three users' tweets would easily be able to discern that users S, K, and L cluster together, while user T is distinct (Fig. 2); additionally, the researcher would be able to make some inferences about how these users differ, which could then be confirmed or further refined by inspecting a representative sample of tweets from each user.

4 Future Work

The results of these two studies serve as a proof-of-concept for two applications of STM to the analysis of social media data: (1) detecting significant events through changes in social media data over time, and (2) grouping and distinguishing authors or sources on the basis of textual content. The first study relied on data taken from before

and after a known major event (the MH 17 crash), while the second study relied on data from authors with relatively clear pro-Russia or pro-Ukraine sympathies. Additional research is needed to determine how STM can be used to detect events in a dataset spanning an arbitrary time range, and to what extent it can be used to detect other author attributes such as gender, age, and nationality.

References

1. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**, 77–84 (2012)
2. Mimno, D.: Computational historiography: data mining in a century of classics journals. J. Comput. Cult. Heritage (JOCCH) **5**, 1–19 (2012)
3. Yang, T.-I., Torget, A.J., Mihalcea, R. Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 96-104. Association for Computational Linguistics, Portland, Oregon (2011)
4. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80–88. Association for Computational Linguistics, New York (2010)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: Working Paper (2014). http://scholar.harvard.edu/bstewart/publications. Accessed 24 September 2014
7. Kumar, S., Barbier, G., Abbasi, M.A., Liu, H.: TweetTracker: an analysis tool for humanitarian and disaster relief. In: Proceedings of the International Conference on Weblogs and Social Media, pp. 661–662. AAAI, California (2011)
8. Kumar, S., Morstatter, F., Liu, H.: Twitter Data Analytics. Springer, New York (2013)
9. Roberts, M.E., Stewart, B.M., Tingley, D.: stm: R Package for Structural Topic Models. Retrieved from The Comprehensive R Network (2014). http://cran.r-project.org/web/packages/stm/stm.pdf
10. New York Times. What Happened to Malaysia Airlines Flight 17. http://www.nytimes.com/interactive/2014/07/18/world/europe/malaysia-airlines-flight-mh17-q-a.html?_r=0 Accessed 23 July 2014
11. Washington Post. A year after Crimean annexation, threat of conflict remains. http://www.washingtonpost.com/world/europe/a-year-after-crimean-annexation-threat-of-conflict-remains/2015/03/18/12e252e6-cd6e-11e4-8730-4f473416e759_story.html Accessed 18 March 2015
12. Crabb, E.S., Mishler, A.M., Paletz, S.B., Hefright, B., Golonka, E.: Reading between the lines: a prototype model for detecting Twitter sockpuppet accounts using language-agnostic processes. Communications in Computer and Information Science (CCIS). Springer, New York (2015)