Towards interactive visualization of public discourse in time and space

Lars Borin
Språkbanken • Dept. of Swedish
University of Gothenburg
Sweden
lars.borin@svenska.gu.se

Tomasz Kosiński
Språkbanken • Dept. of Swedish
University of Gothenburg
Sweden
tomasz.kosinski@gu.se

Abstract

We report on a proof-of-concept study where we (1) apply NLP tools for extracting political-discourse topics from a large Swedish Twitter dataset; and (2) design an interactive spatiotemporal visualization application allowing humanities and social-science scholars to explore how the tweet topics vary over space and time.

1 Introduction

Public discourse has been characterized as being "among the most remarkable inventions of the early 19th century" (Nordmark, 2001, 42). It has been repeatedly transformed over its long history; technologies have evolved, new media have appeared, and participation has become increasingly inclusive. The most recent manifestations of public discourse are the various social media that have emerged only over the last decade or so, complementing or perhaps even supplanting traditional print and broadcast media as the main arena of public discourse and opinion formation, involving many more citizens in a much more interactive mode than ever before.

However, there are many questions about public discourse as conducted in social media, questions about the demography and representativity of participation, whether the issues are the same as in traditional media, and whether public opinion formation processes have become fundamentally different as a result.

Social and political scientists are naturally eager to investigate these and other questions, but face the daunting challenge of dealing with the content of big and streaming textual data. Together with researchers in computer science and language technology they are rising to the challenge (e.g. Conover et al., 2011; Sasahara et al., 2013; Preoţiuc-Pietro et al., 2015). There is still ample scope for methodological development in this area, however, and the work presented below is intended as a contribution to digital humanities and social science methodology. We build on an earlier study of political discussion on Twitter, and, reusing the data from that study, we (a) refine the classification of the content of the tweets using state-of-the-art language processing tools (section 2); and (b) develop an interactive visualization application where the spatiotemporal distribution of the tweet topics along with meta information from the analysis can be explored (section 3).

2 Data and research questions

2.1 Studing political debate on Twitter

The data used for the work presented here comes from an earlier study where Swedish tweets were collected from Twitter's public streaming API during a narrow time window around two televised Swedish party leader debates in October 2013 and May 2014, before the national elections in September 2014.

In the earlier study,[1] basic information retrieval techniques were used to classify the tweets into six topics which had been preselected for the debates, and which are considered to reflect two political issue di-

[1] The study referred to is currently under review for a political-science journal, and due to the double-blind nature of this review process, we are unable to reveal the title and authors of this study here.

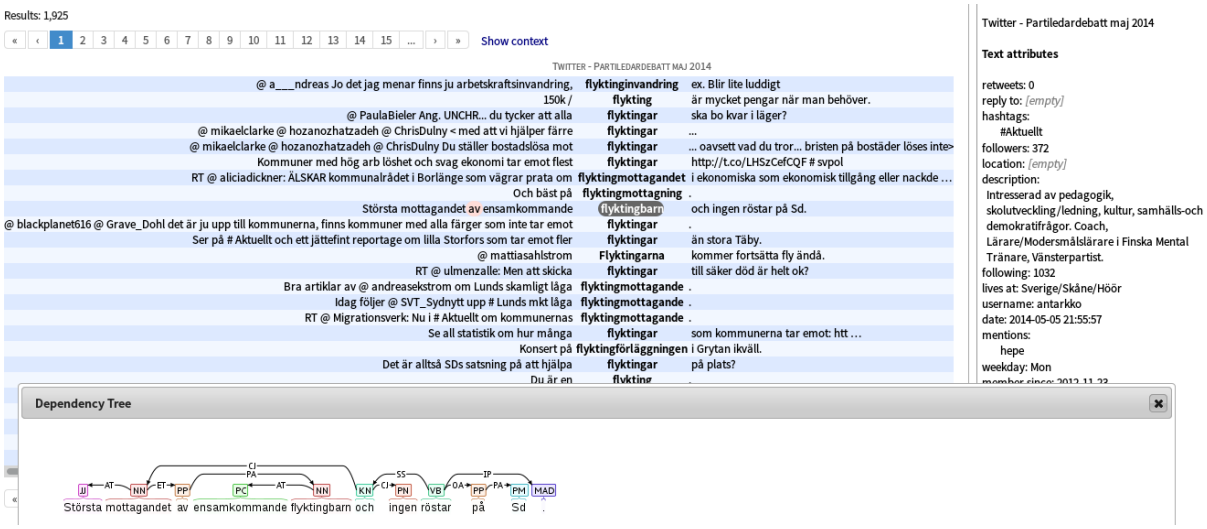

Figure 1: Searching the May 2014 Twitter data using Språkbanken's Korp interface

mensions: *left–right* (topics: *labor market*, *healthcare*, and *education*) and *green/alternative/libertarian–traditional/authoritarian/nationalist* (*GAL–TAN*) (topics: *climate*, *refugees/immigration*, and *crime*).

The tweets were classified into topics using lists of index terms. These lists were incrementally defined by a mixture of manual and automatic methods. Transcriptions of the televised debates and of Swedish parliamentary proceedings formed the basis for the initial, manually constructed, lists. After this, tweets containing at least one word from the initial topic list were merged into a 'topic document' for this particular topic, yielding six multi-tweet topic documents. Additional index terms were identified in these topic documents using the standard tf-idf (term frequency-inverse document frequency) score used for determining index term relevance in information retrieval, and subsequently added to the lists. All list items are text words, and not more abstract linguistic units, such as lemmas or word senses, with the consequence that sometimes several inflected forms of the same lexical item appear in the lists.[2] There are also no multi-word expressions in the lists.

For classification of the tweets, all tweets that contained at least one word from one of the topic lists were considered to discuss the corresponding debate topic. Consequently, tweets could be assigned more than one topic.

The results were presented in numerical form in tables, and additionally in static charts showing pre-elected subsets of tweet frequencies and topic distributions over time.

Some of the research questions addressed in the earlier study relate to the relative frequency of these topics in the tweets (both in relation to each other and in relation to how much time they were accorded in the televised debates), their timing in relation to that in the debates, and whether *GAL–TAN* issues would be more prominent on Twitter than on television, reflecting a hypothesized difference between professional politicians and social-media users.

2.2 Adding natural language processing

The index word lists used for the classification were kindly made available to us by the authors of the earlier study. The datasets used in their study are available through our research unit – *Språkbanken* (the Swedish Language Bank)[3] at the University of Gothenburg – in the form of annotated corpora, containing user and text metadata, (including location and geographical coordinates) and linguistic annotation of the texts: part of speech, lemma, compound segmentation, word sense(s), and dependency syntax, accessible online through our dedicated web interface for interactive corpus queries, called *Korp*,[4] as well as via

[2] Swedish nouns have 8 different inflected forms, verbs have up to 5 forms, and adjectives have maximally 7 forms.

[3] `https://spraakbanken.gu.se/eng`

[4] `https://spraakbanken.gu.se/korp/`. The corpus import pipeline is available for experimentation through a separate web interface at `https://spraakbanken.gu.se/sparv`.

REST web service APIs and as downloadable datasets in sentence-scrambled form (Borin et al., 2012). See Figure 1, illustrating a corpus search for the lemma *flykting* 'refugee', in all its inflected forms and additionally as part of compounds, e.g., the highlighted word *flyktingbarn* 'refugee children'. The NLP tools forming part of Korp's corpus import pipeline are state of the art, but their performance is unequal and heavily dependent on text type, genre, etc. Adesam et al. (2015) describe ongoing work on building an evaluation dataset which will more faithfully reflect the variety of text types and genres found in our corpus collection, and which consequently will allow us to reach a better estimation of the accuracy of the NLP tools that we use for corpus annotation.

The work presented here is part of a larger effort to design e-science tools for research in the humanities and social sciences (HSS) based on massive amounts of text, richly annotated using state-of-the-art language technologies, providing us with a handle on the content of the texts. There are indications that data visualization and visual analytics have an important role to play here (e.g., Havre et al., 2000; Smith, 2002; Schilit and Kolak, 2008; Chuang et al., 2012; Broadwell and Tangherlini, 2012; Krstajić et al., 2012; Sun et al., 2013), and this aspect is the focus of the work presented here.

Thus, we started out by redesigning the earlier study in this direction. The original word lists – containing text word types, i.e., in many cases several inflected forms of the same lexical entry – were run through an automatic morphological analyzer and the output was manually disambiguated. Unanalyzed words were classified into two groups: (1) simplex words missing from the morphological analyzer's lexicon, in many cases typos or irregular spellings; (2) compounds missing from the lexicon, but having received a compound analysis by the morphological analyzer. The first category was left as-is, while the compounds were (manually) reduced to a common prefix or suffix,[5] e.g., *flyktingorganisation* 'refugee organization', *flyktingproblem* 'refugee problem', *flyktingskatastrof* 'refugee disaster', *flyktingsmuggling* 'refugee smuggling', *flyktingstatus* 'refugee status', are all analyzed as compounds with the prefix `flykting..nn.1` 'refugee n'. Hence, we use only the compound prefix as classification criterion.

This resulted in a considerable reduction in the number of index terms. The average number of words per topic in the original study was 219. The average has now been reduced to 161 index terms (a reduction by 26%), but these of course cover many more text word types.

The topic classification now uses the linguistic annotation layers in addition to the text itself, looking for (a) an exact text-word match (i.e., the only classification criterion used in the original study); (b) a lexical entry match; (c) a compound prefix+compound suffix match; (d) a compound prefix match (e.g., `flykting..nn.1` 'refugee n'); or (e) a compound suffix match, in this order of priority. Note that all but the first capture all the inflected forms of a lexicon word, or a maximum of eight forms for a Swedish noun. Note also that matching for compound parts will result in many more compounds being included than those found in the original lists. As in the original study, a tweet may be assigned to multiple classes.

Our classification results are slightly different from those of the earlier study. Notably, the two most common topics – *labor market* and *education* – switch places. This deserves further study, which however falls outside the scope of this presentation.

It has been frequently observed in the literature that the language of social media deviates in various ways from the written standard language, making the use of off-the-shelf NLP tools problematic. We note here that the word lists used in the earlier study contain predominantly orthographically correct items, and the authors of that study also conducted a small manual check, using lemma searches through a corpus search interface, yielding the same proportions of topics as the automatic classification. However, this procedure only gives us an estimation of the *precision* of the classification, but says nothing about its *recall*, which of course is also dependent on how well the NLP tools work with this text type.

In this connection, we note that the morphological analysis used in the present study is quite reliable, building as it does on a full-sized modern Swedish lexical resource (SALDO; see Borin et al., 2013) with about 140,000 entries, covering on the order of two million inflected forms.[6] However, it does not deal with misspellings or with the various manifestations of creative orthography often encountered in social

[5] Here and below, we use "(compound) prefix" and "(compound) suffix" to refer to the first and second member of binary compounds, respectively, i.e., not in the normal linguistic meaning of the terms "prefix" and "suffix".

[6] `https://spraakbanken.gu.se/eng/resource/saldo`

media, so while the precision is predicted to be high also in our case, the recall is – again – unknown. This is clearly something which deserves further, separate, study.

3 Interactive visualization as a research tool for data exploration

Traditional manual text analysis methods founder when faced with so-called big data, e.g., analyzing thousands of newspapers or millions of blog entries. Human limited cognitive capabilities call for help of machines, which don't get tired or bored, also in this case. Contemporary HSS research already leverages possibilities created by automated tools (Grimmer, 2015) and the computational power available today (Lapponi et al., 2013). But in order to benefit of those fully, the challenges posed by the increasing volumes of data generated and collected everyday and frequently made publicly available on request need to be accounted for and addressed. As already mentioned above, an important emerging technology for dealing with very large amounts of textual data is *visual analytics* (Sun et al., 2013). For a number of practical reasons, in our case, a visual text mining application should preferably be accessible through a web interface.[7] Reviewing existing solutions, the following criteria were taken into account:

(C1) Open-source licensing (to be able to make this work publicly available and open);
(C2) support for real-time, interactive visualization of data amounting to millions or billions of records;
(C3) pixel-oriented technique support (Keim, 2000);
(C4) support for the temporal dimension with real-time, interactive browsing;
(C5) user-defined spatial dimension support;[8] and
(C6) support for browsing individual, non-spatiotemporal dimensions independently.

Our tool of choice, which fullfills criteria (C1-C5), is Nanocubes (Lins et al., 2013), an open-source engine for real-time spatiotemporal data exploration. Criterion (C2) makes it possible to analyse corpora consisting of the amount of source data allowing for representative analysis of textual data sources. Criterion (C3) refers to the relevance of pixel-oriented technique for large spatial datasets visual exploration tools. Criterion (C4) allows for more focused visual search, analysing only a selected time frame at a time, and makes it easier to structure. Criterion (C5) makes it possible to provide non-spatial datasets with a self-designed, simulated spatial domain, supplementing the dataset with a new meaning, integrated with the existing visualization feature, i.e. dimensionality. Criterion (C6) was fullfilled by extending Nanocube's frontend within the presented work.

Visualizing data in two-dimensional space implicitly reduces cognitive load for the user as at least two pieces of information, e.g., latitude and longitude, are presented in the familiar way. All of those features enable real-time sense-making with reproducibility of performed searches, while the user has permanent access to the complete real-world dataset underlying the visualization (Baker et al., 2009).

Using Nanocubes as the data visualization engine, we have established that it is possible to browse information derived from over 20 million Swedish tweets across not only the spatial and temporal dimensions, but also at least 8 other, user-defined dimensions in a highly interactive way. Nanocubes aggregates the data for efficiency and provides no 'way back' to the original data. However, since we believe that this type of visualization will be acceptable to HSS researchers only if they can also at all times inspect the underlying textual data, we have extended Nanocubes with a lookup feature addressing this need. The mechanism behind this feature takes advantage of visual browsing performed by the user with the use of on-screen controls, e.g. buttons, drop-down menus or regions drawn on the top layer of the visualization as well as panning and zooming. Then the user, with each step narrowing down the selected spatial, temporal or categorical dimension, implicitly constructs a corpus query translated into criteria narrowing down the subset of all visualized records. When the user selects the 'dive in' option, he or she is presented with a corpus view of the subset of source material selected based on visual browsing

[7] The reasons for preferring a web interface are not restricted to visual analytics applications, but apply to all kinds of interactive interfaces presenting the results of processing large datasets. A web interface can draw on the generally larger processing and storage capacity of (clusters of) servers, as compared to desktop or laptop computers, so that users can access and process large textual datasets without the need for their own machine to have high-performance or large-storage capabilities. A web interface can further be kept up to date by making changes in one place only, and – quite crucial in many university settings – users will not be dependent on having the administrative privileges required to install client software.

[8] This could be, e.g., a two-dimensional projection of a multidimensional document vector space model.

criteria specified before.

For the proof-of-concept study described here, we used the Swedish Twitter data described above in section 2, providing it with three user-defined dimensions: *Topic* (described above), *Type of match* (which kind of index match was found) and *Strength of evidence* (how many matching words were found). It is evident that this visualization provides added functionality in comparison to the earlier study. Notably, we can explore whether the topic proportions in the tweets are different in different parts of the country (which they seem to be to some extent), selecting moments of interest of the debate. Depending on the resolution of the underlying geolocation data, we can zoom in to even see whether city neighborhoods behave differently with respect to the investigated variables. All the visualizations are available online as a part of bigger work in progress developed to address HSS needs as mentioned before (see Figure 2).

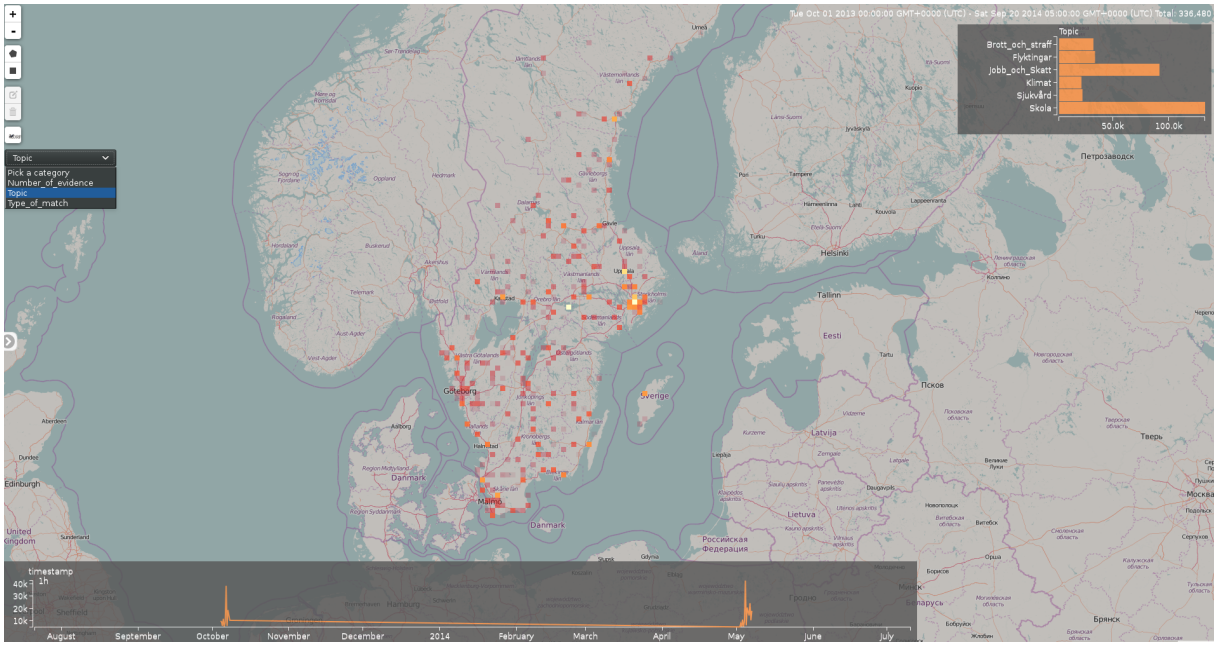

Figure 2: Interactive visualization of the Swedish Twitter debate topics with Nanocubes

There are many issues that remain unaddressed in our small proof-of-concept study. For instance, only about 35% of the tweets could be geolocated. For this we used two kinds of metadata: (1) explicit geographical coordinates, provided in about 17% of all tweets; and (2) matching of words in the "location" metadata against a gazetteer of Swedish place names downloaded from the Swedish postal services, which yielded an additional 18%. Clearly, it would be desirable to do better, perhaps using methods similar to those suggested by Berggren et al. (2015), who geolocate Swedish tweets based on regionally characteristic vocabulary automatically inferred from tweets with explicit location information (mainly proper nouns, but also some dialectal words).

4 Conclusions and future work

We have presented a proof-of-concept interactive spatiotemporal visualization of the results of processing a large Twitter dataset with state-of-the-art NLP tools, enabling more detailed and varied exploration of the research questions of the original study for which the data were collected.

There are several directions in which we intend to continue this work. We think it could be rewarding to enter into a collaboration with the authors of the previous study to explore the usefulness of the kind of spatiotemporal visualization discussed here, as well as investigate the influence on the classification of the NLP tools used. As mentioned above, it is desirable to be able to geolocate more than about a third of the tweets. Also, in order to automate the data pre-processing phase and enable users to visually and interactively analyse the dataset of their choice, the existing visualization engine needs to be integrated with a tool allowing for data preprocessing and formatting, without a limit to the maximal number of

records which can be processed.

Other kinds of automated NLP classification will also be added to the datasets as they become available in the corpus import pipeline, e.g., multi-word expressions, word senses, sentiment and argumentation analysis, as well as other methods for topic classification (e.g., LDA or HDP topic modelling), which will help us to throw more light on questions of political opinion formation and expression in social media.

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*;[9] contract 2012-5738) as well as funding to Swedish CLARIN (*Swe-Clarin*;[10] contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

References

Yvonne Adesam, Gerlof Bouma, and Richard Johansson. 2015. Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 1–9, Vilnius. NEALT.

Jeff Baker, Jim Burkman, and Donald R. Jones. 2009. Using visual representations of data to enhance sensemaking in data exploration tasks. *Journal of the Association of Information Systems*, 10(7):533–559.

Max Berggren, Jussi Karlgren, Robert Östling, and Mikael Parkvall. 2015. Inferring the location of authors from words in their texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 211–218, Vilnius. NEALT.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Peter M. Broadwell and Timothy R. Tangherlini. 2012. TrollFinder: Geo-semantic exploration of a very large corpus of Danish folklore. In *The Third Workshop on Computational Models of Narrative*, pages 50–57, Istanbul. ELRA.

Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 89–96, Barcelona. AAAI.

Justin Grimmer. 2015. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48:80–83, 1.

Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, Salt Lake City.

Daniel A. Keim. 2000. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, January.

Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A. Keim. 2012. Incremental visual text analytics of news story development. In *Conference on Visualization and Data Analysis (VDA '12)*.

Emanuele Lapponi, Erik Velldal, Nikolay Vasov, and Stephan Oepen. 2013. HPC-ready language analysis for human beings. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 447–452, Oslo. NEALT.

Lauro Lins, James T. Klosowski, and Carlos Scheidegger. 2013. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, Dec.

[9] `https://spraakbanken.gu.se/eng/culturomics`

[10] `https://sweclarin.se/eng`

Dag Nordmark. 2001. Liberalernas segertåg (1830–1858). In Karl-Erik Gustafsson and Per Rydén, editors, *Den svenska pressens historia. II: Åren då allting hände (1830–1897)*, pages 18–125. Ekerlids förlag, Stockholm.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of ACL 2015 (Volume 1: Long Papers)*, pages 1754–1764. ACL.

Kazutoshi Sasahara, Yoshito Hirata, Masashi Toyoda, Masaru Kitsuregawa, and Kazuyuki Aihara. 2013. Quantifying collective attention from Tweet stream. *PLOS ONE*, 8(4):e61823.

Bill N. Schilit and Okan Kolak. 2008. Exploring a digital library through key ideas. In *Proceedings of JCDL'08*, pages 177–186, Pittsburgh. ACM.

David A. Smith. 2002. Detecting and browsing events in unstructured text. In *SIGIR'02*, Tampere. ACM.

Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.