An Unsupervised System for Visual Exploration of Twitter Conversations

Derrick Higgins, Michael Heilman, Adrianna Jelesnianska and **Keith Ingersoll**
Civis Analytics
dhiggins@civisanalytics.com

Abstract

Social media provides a wealth of information regarding users' perspectives on issues, public figures and brands, but it can be a time-consuming and labor-intensive process to develop data pipelines in which those perspectives are encoded, and to build visualizations that illuminate important developments. This paper describes a system for quickly developing a model of the conversation around an issue on Twitter, and a flexible visualization system that allows analysts to interactively explore key facets of the analysis.

1 Introduction

This paper introduces a visualization system designed to support deep analysis of Twitter conversations that persist over a long period of time (as well as inform decisions concerned with recent developments). Given a set of tweets that define the conversation to be modeled (and which may be selected by arbitrary criteria including keywords, users, and time frames of interest), Civis Analytics' web-based system provides a rich and interactive set of visualizations that illustrate

- the major themes of discussion and user perspectives,
- communities of users engaged in the conversation (and their centrality to it),
- how strongly users and user groups address particular themes,
- the activity of themes and user groups over time, and
- the activity of themes and user groups in different geographical regions.

We anticipate that an analyst can use this tool to quickly develop a high-level understanding of the dynamics of a given issue on Twitter, to drill down into the specifics of how particular users and tweets contribute to changes in the conversation, and to develop informed strategies for outreach or intervention. The tool takes advantage of current research in topic modeling and community analysis, and can easily incorporate domain-specific knowledge where it is available.

2 Related Work

Civis' system for visualizing the dynamics of social media conversations builds upon prior research in a number of areas.

2.1 Data processing technologies

First, there are a number of enabling technologies that are used to transform the unstructured data stream from Twitter into a structured representation that our system can use as the basis for visualization.

One such technology is topic modeling – unsupervised identification of major themes in a set of text documents. Since its introduction (Blei et al., 2003), topic modeling has been tailored to perform better on short texts such as microblogs. For example, the Dirichlet Multinomial Mixture Model (Yin and Wang, 2014) modifies the standard LDA model to constrain all words of a text to be generated by

Proceedings of NAACL-HLT 2016, pages 60–65,
San Diego, California, June 12-17, 2016.

the same mixture component (topic), and the Biterm Topic Model (Yan et al., 2015) combines adjacent terms to provide richer information about the subject of a short text. The topic modeling method incorporated into the Civis social listening tool builds upon this research to provide meaningful topic groupings on short texts, but currently does not make use of any tweet metadata such as author or time frame.

Similarly, much progress has been made in recent years in adapting sentiment analysis to microblog data. Promising methods include hybrid topic-sentiment models (Xiang and Zhou, 2014), the incorporation of tweet metadata into the model (Vosoughi et al., 2015), and the application of word sense disambiguation as a pre-processing step (Sumanth and Inkpen, 2015).

Finally, our visualization tool makes use of graph-based community analysis. We use the Louvain algorithm (Blondel et al., 2008) on the friend-follower graph to find coherent user communities in a Twitter conversation; a survey of modern methods for community analysis is given by Malliaros and Vazirgiannis (2013).

While the current version of the Civis visualization tool includes topic modeling, community inference, and sentiment analysis, it is agnostic as to the source of these data layers. Our data abstraction layer guarantees that the visualization too will work as long as the relevant model can assign numeric values to tweets (or, in the case of community analysis, categories to users).

2.2 Visualization research

We also build on previous work that has developed interfaces for visualizing microblog metadata derived using natural language tools. Xie et al. (2013) use high-frequency bar charts to demonstrate the activity associated with "bursty" topics identified by their algorithm, while Malik et al. (2013) conduct user studies to refine a Twitter topic exploration tool with multiple information displays.

In the area of sentiment analysis, many dashboards have been developed to show the time course of opinion associated with a particular topic or keyword. Hao et al. (2011) develop a system with more engaging displays than the standard line chart, including heatmaps embedded within calendars, and activity maps.

The tool we will present in our demonstration differs from these previous interfaces primarily in the flexibility it allows for faceting the analysis according to NLP-related features, location, user community, and time.

3 System Overview

Civis Analytics' system for visualizing conversations on Twitter is intended to allow an analyst to quickly develop a model of the most important themes of discussion on an issue, the parties involved in the conversation, and trends across time and space.

In our live demonstration of the system, we will show how it can be used to analyze a wide variety of issues, including climate change, data science, and education. To provide a clearer indication of the type of views the system supports, though, we focus here on an analysis of the 2015-16 US Presidential primary campaign. The dataset for our analysis consists of approximately 15 million tweets posted between May, 2015 and February, 2016. These tweets were selected to include references to one or more of the Republican or Democratic candidates for the presidency (by name or using a strongly associated hashtag).

We applied a version of topic modeling adapted to short texts to infer a set of 44 topics, which are not mutually exclusive. Some of these topics are associated with particular candidates, while others are related to issues or events in the campaign. An analyst manually labeled these topics by inspecting the words and tweets most strongly associated with each, and the resulting topic definitions are included in our visualization tool for reference; Figure 1 demonstrates how topic-word associations are shown to users for the first 5 topics in our set (alphabetically).

We also selected a set of the most-active and most-followed users in the conversation for community analysis. We automatically inferred a set of three communities of users, which roughly correspond to political orientation ("Conservatives", "Progressives", and "Media", with the last group also including some political centrists). In our visual interface, these user groupings can be explored using an interactive graph that shows user clusters

projected into two dimensions (cf. Figure 2). The 2D projection is done using the t-SNE algorithm (van der Maaten and Hinton, 2008) so that users with similar friend-follower relationships generally appear near one another. Users can interactively explore the graph to identify key users in each community and access their profiles.

The Civis interface also allows users to access more information about particular Twitter accounts that are relatively central to the network for a particular issue (using a network centrality measure such as PageRank). We hypothesize that accounts with a central position in the friend-follower network restricted to a particular issue are likely to be important, and perhaps influential figures in the relevant Twitter community. As shown in Figure 3, the available metadata about these key accounts includes the number of followers, how active they are on a given issue, and their relative engagement across topics.

Our system also supports sentiment analysis; Figure 4 demonstrates the average sentiment polarity of tweets overlaid on the graph of relationships between Twitter users. While the activity of particular topics is often strongly indicative of sentiment on its own, the sentiment layer may sometimes encode independent information as well.

Figure 4: Overlay of sentiment on to user exploration view

Finally, given base data associated with each tweet regarding topics, sentiment, and user communities, the Civis social listening tool provides multiple visualizations that allow users to explore the inter-relationships among these variables (including crosstabs, maps, and time series charts). Figure 5 shows a sample line chart displaying the time course of activation of the Bernie Sanders, Hillary Clinton and Black Lives Matter topics, restricted to the subset of Twitter users who are classified as belonging to the Progressive community.

4 Discussion and Future Work

The Civis social listening tool provides a flexible environment in which an analyst can interactively explore selected dimensions of a Twitter conversation. We hope it will inspire future research on visualization techniques for social media, and we anticipate extending the tool in a number of directions in the near future.

First, we plan to extend the set of case studies we have developed in order to test and demonstrate its functionality. We have built out analyses of the 2016 US Presidential Primary, climate change, and data science, but we would like to extend this set to include a more diverse set of topic areas.

Second, we plan to conduct more formal usability studies to establish which parts of the interface are most useful, and which users may find confusing or superfluous.

Finally, we hope to broaden the source data behind the tool to include other social media platforms and text sources. Print news stories, blog posts, and data feeds from Instagram all represent extensions that could provide additional context for analysts to exploit, given the right tools for summarization of their content.

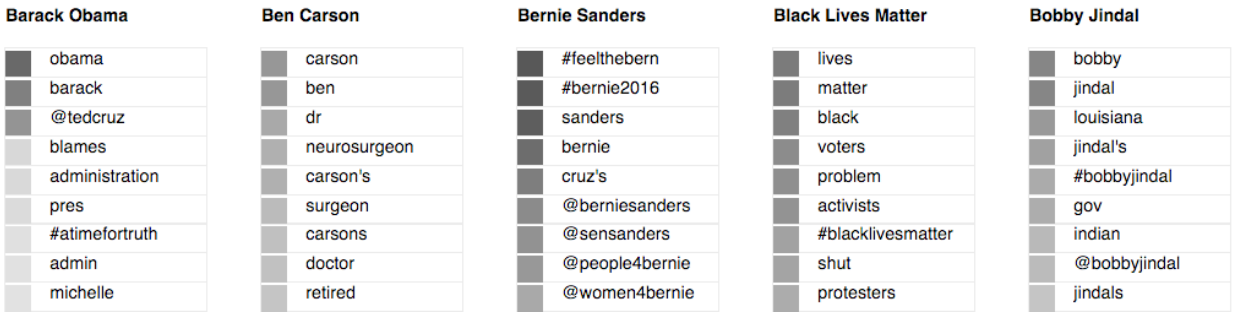

Figure 1: Visualization of words most strongly associated with selected topics

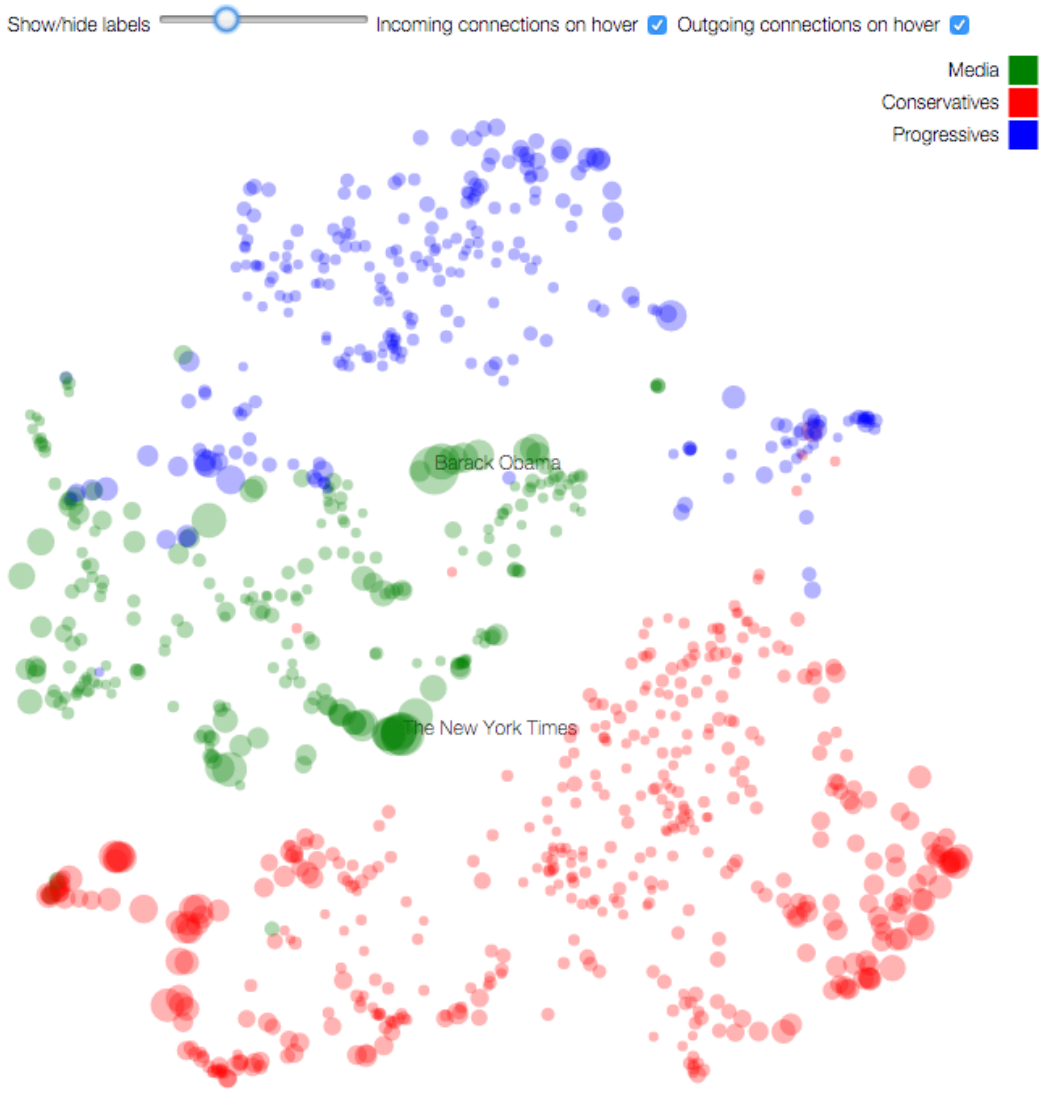

Figure 2: Interface for exploration of users and communities engaged in Twitter conversation (using a two-dimensional projection of friend-follower links)

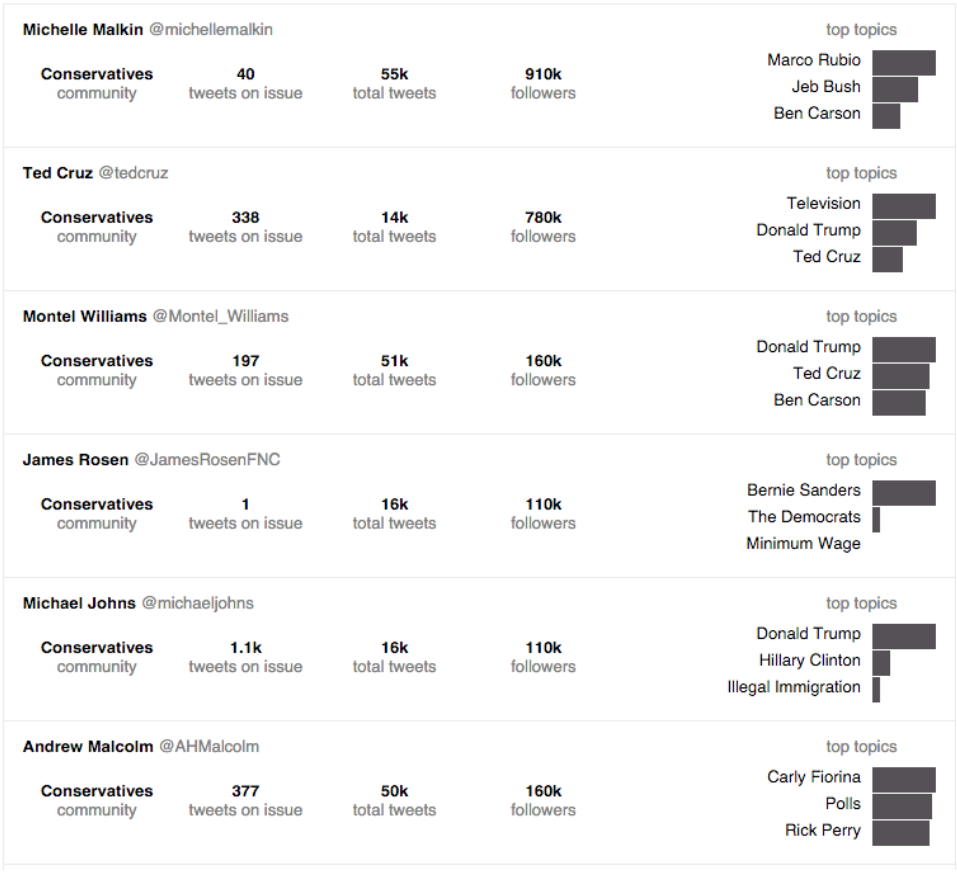

Figure 3: List of most influential accounts from the Conservative community in the Presidential Primary analysis

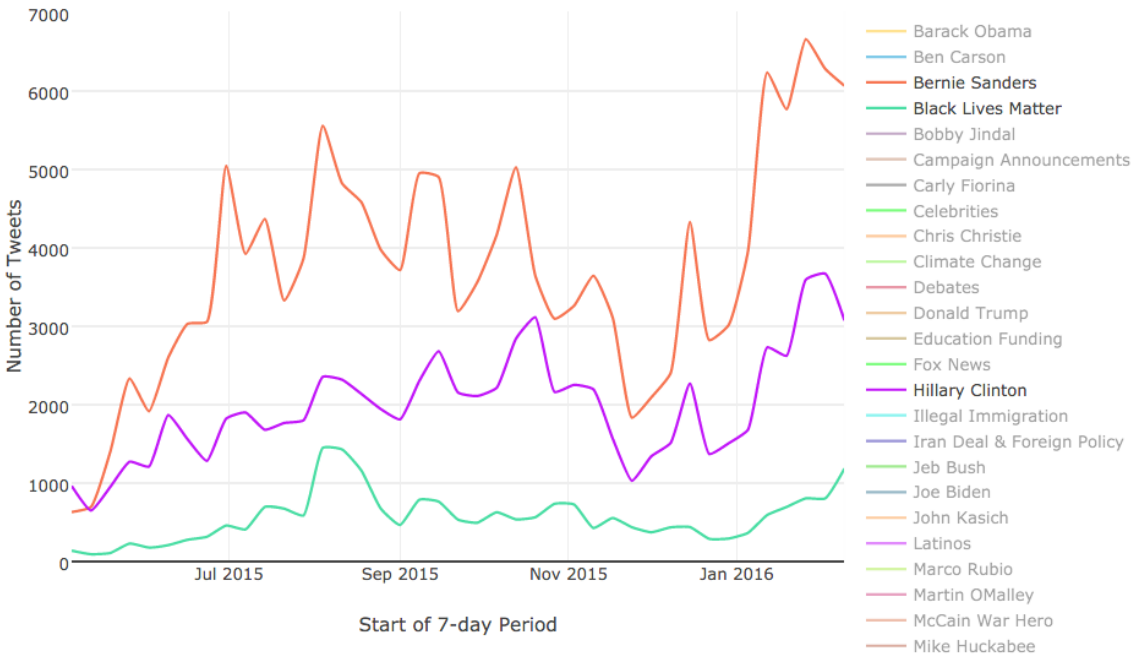

Figure 5: Visualization of trends in the Civis social listening interface, illustrated using the activity of Progressive users on the Bernie Sanders (top), Hillary Clinton (middle) and Black Lives Matter (bottom) topics

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):10008+, July.

Ming C. Hao, Christian Rohrdantz, Halldor Janetzko, Umeshwar Dayal, Daniel A. Keim, Lars-Erik Haug, and Meichun Hsu. 2011. Visual sentiment analysis on Twitter data streams. In *IEEE VAST*, pages 277–278. IEEE Computer Society.

Sana Malik, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben Shneiderman. 2013. TopicFlow: visualizing topic alignment of Twitter data over time. In Jon G. Rokne and Christos Faloutsos, editors, *ASONAM*, pages 720–726. ACM.

Fragkiskos D. Malliaros and Michalis Vazirgiannis. 2013. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95 – 142. Clustering and Community Detection in Directed Networks: A Survey.

Chiraag Sumanth and Diana Inkpen. 2015. How much does word sense disambiguation help in sentiment analysis of micropost data? In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 115–121, Lisboa, Portugal, September. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Soroush Vosoughi, Helen Zhou, and deb roy. 2015. Enhanced Twitter sentiment classification using contextual information. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 16–24, Lisboa, Portugal, September. Association for Computational Linguistics.

Bing Xiang and Liang Zhou. 2014. Improving Twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 434–439, Baltimore, Maryland, June. Association for Computational Linguistics.

Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. 2013. TopicSketch: Real-time bursty topic detection from twitter. In Hui Xiong, George Karypis, Bhavani M. Thuraisingham, Diane J. Cook, and Xindong Wu, editors, *ICDM*, pages 837–846. IEEE Computer Society.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. A probabilistic model for bursty topic discovery in microblogs. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jianhua Yin and Jianyong Wang. 2014. A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 233–242, New York, NY, USA. ACM.