TweetVista: An AI-Powered Interactive Tool for Exploring Conversations on Twitter

Prashanth Vijayaraghavan
MIT Media Lab
Cambridge, MA 02139, USA
pralav@mit.edu

Soroush Vosoughi
MIT Media Lab
Cambridge, MA 02139, USA
soroush@mit.edu

Ann Yuan
MIT Media Lab
Cambridge, MA 02139, USA
annyuan@gmail.com

Deb Roy
MIT Media Lab
Cambridge, MA 02139, USA
dkroy@mit.edu

IUI'17 Companion, March 13-16, 2017, Limassol, Cyprus
ACM 978-1-4503-4893-5/17/03.
http://dx.doi.org/10.1145/3030024.3040979

Abstract
We present *TweetVista*, an interactive web-based tool for mapping the conversation landscapes on Twitter. TweetVista is an intelligent and interactive desktop web application for exploring the conversation landscapes on Twitter. Given a dataset of tweets, the tool uses advanced NLP techniques using deep neural networks and a scalable clustering algorithm to map out coherent conversation clusters. The interactive visualization engine then enables the users to explore these clusters. We ran three case studies using datasets about the 2016 US presidential election and the summer 2016 Orlando shooting. Despite the enormous size of these datasets, using TweetVista users were able to quickly and clearly make sense of the various conversation topics around these datasets.

Author Keywords
Twitter; Tweet2Vec; Conversation Clusters; Semantic Clusters; Interactive Tool

ACM Classification Keywords
H.5.m [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous; H.3.3 [Information Search and Retrieval]: Clustering

Introduction

Twitter should be an ideal place to get a fresh read on how different issues are playing with the public, one that's potentially more reflective of democracy in this new media age than traditional polls. Pollsters typically ask people a fix set of questions, while in social media people use their own voices to speak about whatever is on their minds. Millions are discussing politics and other issues every day on Facebook, Twitter, Instagram and other platforms, right alongside the candidates and the journalists covering them.

However, the sheer scale of the data on Twitter presents both opportunities and challenges. On the plus side, this allows us to measure in aggregate how the various issues are rising and falling in prominence over time. But unlike traditional news coverage, which is based on journalistic observation, interviews and storytelling, data is about numbers. It lacks the human voices and faces that make for compelling stories. For years, media outlets have been working around this problem by paying attention to the most popular conversation on social media, often identified through hashtags, and anecdotally pulling out citizen comments on those topics. The problem with trending topics is that can overlook non-viral issues that many people care about. And anecdotally selected tweets are not necessarily reflective of the larger conversation.

Using recent advances in deep neural networks for natural language processing, we developed a way to automatically identify various clusters of any conversation on Twitter. We also built a web-based interactive tool, called *TweetVista*, on top of our system that allows users to explore the conversation landscapes identified by our system. The following sections contain detailed descriptions of various components of the tool, implementation details of the tool and case studies.

TweetVista

TweetVista is composed of three parts: (a) a sophisticated mechanism for extract rich semantic features from the tweet text, (b) a scalable methodology to agglomerate semantically similar tweets into a cluster, (c) an interactive endpoint to visualize the tweet clusters. Below we explain each of these three sections in detail.

Tweet2Vec

Due to the noisy nature of tweets, commonly used methods to extract semantic features such as TF-IDF [6] and distributed word vectors [4], operating at word-level, do not perform well. Therefore, we utilized Tweet2Vec [7], a character-level CNN-LSTM encoder-decoder approach, to learn general purpose vector representation of tweets. These vectors capture abstract semantic structures that can be applied to several generic tasks. Tweet2Vec [7] is a recent method for generating general-purpose vector representation of tweets. Tweet2Vec removes the need for expansive feature engineering and can be used to train any standard off-the-shelf classifier (e.g., logistic regression, svm, etc). It uses a CNN-LSTM encoder-decoder model that operates at the character level and can deal with the noise and idiosyncrasies in tweets. Character-level models are great for noisy and unstructured text since they are robust to errors and misspellings in the text. The model learns abstract textual concepts from the character level input of tweets.

We trained our model on 5 million randomly selected English-language tweets populated using data augmentation techniques, which are useful for controlling generalization error for deep learning models. Data augmentation involved replacing some of the words with their synonyms as mentioned in [8, 7].

Case Studies
We tested TweetVista on three different datasets: 1) Trump's Immigration Speech, 2015, 2) Orlando Shooting & aftermath, 2016, 3) Discussion of US Economy on Twitter, Summer 2016.

These datasets were collected using a state-of-the-art supervised Twitter topic classifier [5]. (The details of the topic classifier are out of the scope of this paper, please read the cited paper for more details.)

Figure 2 shows a 2D view of the conversation clusters generated by TweetVista, for the "US economy" dataset. In the interest of brevity we do not show the clusters for all the datasets but all three datasets can be fully explored in 2D and 3D using our tool on TweetVista.com. As can be seen, an interesting narrative emerges from these tweet clusters.

Clustering
Next, we cluster the tweets based on the tweet embeddings generated by Tweet2Vec (with a vector size of 256) to aggregate semantically similar tweets into a topic bucket. This requires a scalable clustering technique that can take a large number of tweets as input and cluster them in a non-parameterized setting. We used a scalable, non-parameterized hierarchical density-based clustering algorithm called Hierarchical DBSCAN (HDBSCAN), introduced by Campello, et.al[1]. HDBSCAN, is a clustering algorithm that can be seen as an improvement over existing density-based clustering algorithms.

Visualization Engine
Finally, the visualization engine renders clusters of semantically related tweets as a particle cloud. Users can explore tweets by panning, rotating, or zooming the cloud. Users can filter the tweets shown by properties of their content or authors. The interface also includes details regarding each semantic cluster. Users can choose between several different 2D and 3D datasets to visualize using the tool.

The first step for visualization is the reduction of the high dimensional tweet embeddings to two or three dimensions. We used *t-SNE* for this task [3]. t-SNE is a variation of Stochastic Neighbor Embedding [2] that is easier to optimize, and produces significantly better visualizations by minimizing the tendency to crowd points together in the center of the map.

Interface Overview
TweetVista is a desktop web application best viewed with Google Chrome. A short video, fully describing the interface can be found at https://youtu.be/BBtZ6P4FLds. Figure 1 shows a screenshot of TweetVista's interface. Upon loading TweetVista, users see a visualization of the default dataset. Users can visualize other datasets by making a selection

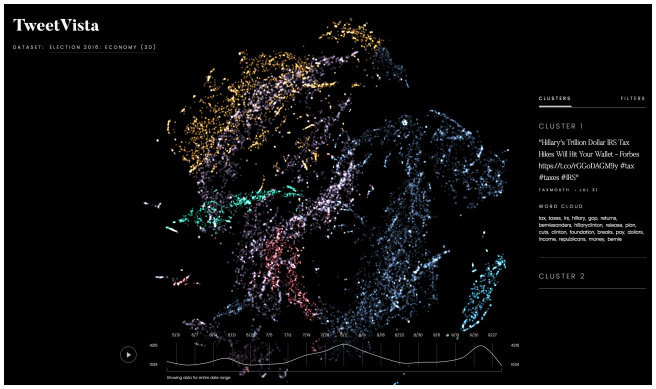

Figure 1: Screenshot of the entire interface of TweetVista.

within a pop-up menu that appears when they click on the title of the interface. Tweets are represented by particles whose position in 3D is determined by the t-SNE algorithm described earlier. Users can zoom into the cloud of tweets by using their mouse wheel or trackpad, they can rotate it by dragging along the interface, and they can also pan the cloud's position by pressing their arrow keys. Tweet particles are colored according to the conversation cluster they belong to. Users can see the text, author, and date of each tweet by hovering over it. Users can filter the tweets shown by content properties such as the civility of the tweet (e.g., whether the tweet contains profanity), or properties of the author such as whether the account is verified, the author's number of followers, statuses, followees, etc. Users can also filter the date range from which data is drawn by manipulating the timeline at the bottom of the interface. They can select a particular pre-defined date interval, or press the "play" icon, which allows them to see the clusters forming over time. The interface also includes a side-panel that provides details about each semantic cluster of tweets

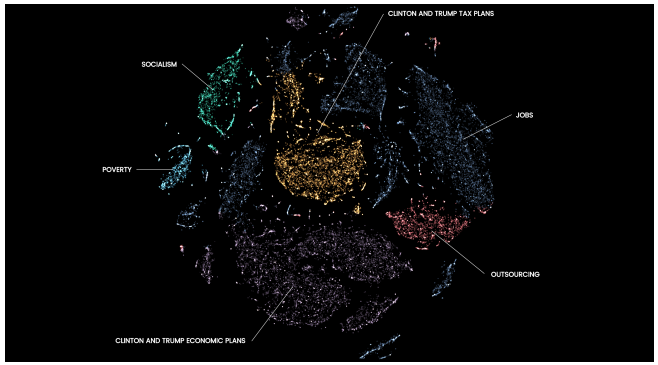

Figure 2: Conversation clusters around the topic of the US economy in the context of 2016 presidential election.

(such as the most frequently occurring words in the cluster) in the currently visualized dataset.

Conclusion

In this paper, we presented TweetVista, an interactive tool for mapping the conversation landscapes on Twitter, to better understand what the Twitter public is saying about various issues. TweetVista has two main components: 1) Identification of conversation clusters for a given dataset, 2) An interactive visualization enabling users to explore the landscape of tweet conversations for a given topic. We tested TweetVista on three datasets from 2015 and 2016.

TweetVista has at its core, a powerful semantic analysis engine that utilizes recent advances in natural language processing using deep neural networks. In contrast to similar tools, our tool was specifically designed to deal with the short, noisy and idiosyncratic nature of tweets. TweetVista enables users to make sense of large volumes of tweets about any given topic.

REFERENCES

1. Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 160–172.
2. Geoffrey E Hinton and Sam T Roweis. 2002. Stochastic neighbor embedding. In *Advances in neural information processing systems*. 833–840.
3. Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.
4. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
5. Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. 2016. Automatic Detection and Categorization of Election-Related Tweets. In *Proceedings of the Tenth ICWSM*.
6. Soroush Vosoughi and Deb Roy. 2016. A Semi-Automatic Method for Efficient Detection of Stories on Social Media. In *Proceedings of the Tenth ICWSM*.
7. Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. 2016. Tweet2Vec: Learning Tweet Embeddings using Character-level CNN-LSTM Encoder-Decoder. In *Proceedings of the 39th SIGIR*.
8. Xiang Zhang and Yann LeCun. 2015. Text Understanding from Scratch. *arXiv preprint arXiv:1502.01710* (2015).