# The Social Media Macroscope

www.socialmediamacroscope.org


Joseph T. Yun
*University of Illinois*
*Technology Services*
Urbana, Il, USA
jtyun@illinois.edu

Nickolas Vance
*University of Illinois*
*Technology Services*
Urbana, Il, USA
npvance2@illinois.edu

Chen Wang
*University of Illinois*
*National Center for*
*Supercomputing Applications*
Urbana, Il, USA
cwang138@illinois.edu

Joseph Troy
*University of Illinois*
*Library*
Urbana, Il, USA
jmtroy@illinois.edu

Luigi Marini
*University of Illinois*
*National Center for*
*Supercomputing Applications*
Urbana, Il, USA
lmarini@illinois.edu

Robert Booth
*University of Illinois*
*Technology Services*
Urbana, Il, USA
booth@illinois.edu

Todd Nelson
*University of Illinois*
*Technology Services*
Urbana, Il, USA
tjn@illinois.edu

Ashley Hetrick
*University of Illinois*
*Library*
Urbana, Il, USA
ahetrick@illinois.edu

Hagen Hodgekins
*Elizabeth State University*
Elizabeth City, NC, USA
hagen.hodgkins@gmail.com



***Abstract*— In recent years, the explosion of social media platforms and the public collection of social data has brought forth a growing desire and need for research capabilities in the realm of social media and social data analytics. Research on this scale, however, requires a high level of computational and data-science expertise, limiting the researchers who are capable of undertaking social media data-driven research to those with significant computational expertise or those who have access to such experts as part of their research team. The Social Media Macroscope (SMM) is a science gateway with the goal of removing that limitation and making social media data, analytics, and visualization tools accessible to researchers and students of all levels of expertise. The SMM provides a single point of access to a suite of intuitive web interfaces for performing social media data collection, analysis, and visualization via for open-source and commercial tools. Within the SMM social scientists are able to process and store large datasets and collaborate with other researchers by sharing ideas, data, and methods. This document functions as a brief primer on the initial build of the SMM.**


## I. Introduction

Social media analytics is, "concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application" [1, p. 14]. The global social media analytics market was valued at $3.07 billion in 2017 and is projected to grow to $16.37 billion by 2023 [2]. The analysis of social media is not limited just to business applications, but it is also used widely in academic research, *e.g.*, [3] and [4]. One of the barriers to entry with regards to social media analytics is that conducting social media analytics requires an understanding of data science principles as well as a range of skills required to conduct these data science methods [5]. The SMM helps to bridge this gap by providing intuitive web interfaces for conducting social media analytics. We describe the SMM in greater detail within this paper by focusing on the first tool that we have made available within the SMM, namely the Social Media Intelligence and Learning Environment (SMILE). Before we look deeper into SMILE, we provide a brief inventory of social media analytics tools previously made available to researchers, and how the SMM addresses gaps that exist with these tools.

## II. Background

There is no shortage of social media analytics tools and environments across both industry and academia, but most of these tools and environments are built in a way that favors usage by companies (as opposed to academia), *e.g.*, [6], or those with computer science backgrounds, *e.g.*, [7]. The gap therefore exists for academic researchers that do not necessarily come from computational backgrounds. To illustrate how the SMM fills this gap, we present two exemplar researchers: a researcher without a background in the computational sciences that wants to use social media analytics to answer social research questions (SocRes), and a researcher with a background in the computational sciences that wants to build data science models/methods for social media analytics (CompRes). As a note, we acknowledge that researchers are evolving to be somewhere in the middle between the two exemplars that we present, but we present our argument in this manner to clarify the gap in which the SMM addresses.


*Presented at Gateways 2018, University of Texas, Austin, TX, September 25–27, 2018.*
*https://gateways2018.figshare.com/*

If we were to consider what SocRes may want from a social media analytics research environment, two major requirements may be the desire for the environment to be easy to use and to have models that are transparent in their methodology. These two axes are pictured in Fig. 1 along with the placement of some common tools/environments used for social media analytics.

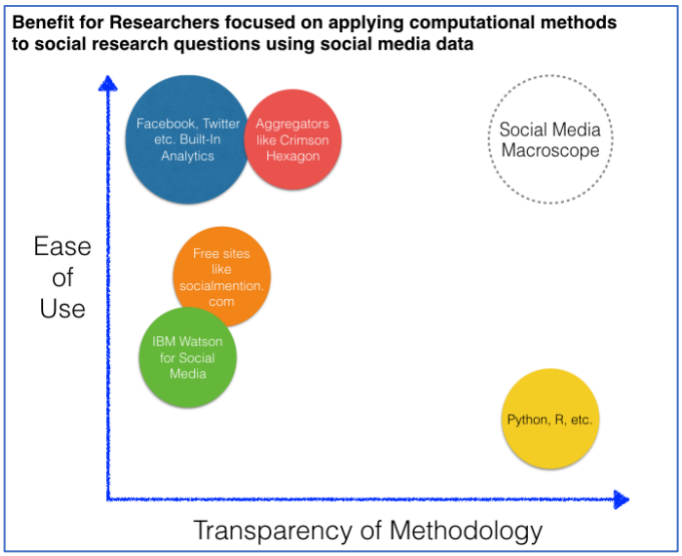


Fig. 1. The value of the Social Media Macroscope for social researchers.

As can be seen in Fig. 1, there are many tools for social media analytics that are easy to use, such as Crimson Hexagon [6], or the built-in analytics tools for Facebook of Twitter. The problem is that their tools commonly are "black-box" in that a researcher will not know exactly how the data science models are specifically built, thus the methodology is not transparent (e.g., there may be sentiment analysis without an explanation of how they are detecting sentiment). The opposite example of fully transparent methodology but hard to use is directly using programming languages such as Python or R (with the associated packages for data science) to analyze social media data.

Now if we consider what CompRes might want from a social media analytics environment, their values are potentially different as shown in Fig. 2. Since CompRes most likely already knows how to program and use the data science packages available for Python and R, they may be more concerned with getting access to more social media data than is available via free public APIs. Social media data aggregators, such as Crimson Hexagon, provide much more data than what is available via free public APIs [8], but they are expensive to subscribe to. The SMM will house a large repository of social media data over time, thus benefitting CompRes. CompRes also may consider the need to have a place where they can house their novel social media analytics models that is more readily accessible to SocRes. Since CompRes' objectives as an academic researcher is not to provide a polished and easy to use interface to their model, their models usually end up being published only in two forms: a conference paper, *e.g.*, [7], and code deposited to a GitHub repository, *e.g.*, [9]. This poses a problem for SocRes who may not have the background to take the code from the GitHub repository and use it to answer their social research question. Thus, the SMM aims to take these types of open code packages and build them into an intuitive web interface (with the desire to partner with CompRes to help build their code into the SMM).

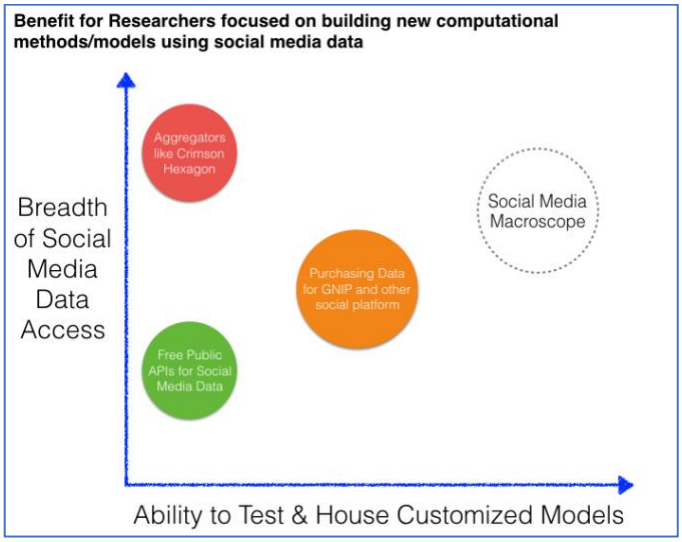


Fig. 2. The value of the Social Media Macroscope for computational researchers.

Now there are quite a few social media analytics tools used in academic research and we will mention a couple to illustrate the differences between them and the SMM. With this said, there is no technical limitation for these tools to be offered via the SMM as will become apparent in the architecture section of this paper.

One of the most popular social media analytics tools used in research is NodeXL [10]. NodeXL is a plug-in to Microsoft Excel that provides a researcher the ability to easily download social media data, apply network analysis to the data, and conduct some forms of dictionary-based analysis on the data. Although an exceptional tool, some of the limitations is that NodeXL only works on Windows PCs (or virtual installations of Windows) and it is largely built only for network analysis. In contrast, the SMM is a web-based environment that can be accessed by any computer with a web browser, and the SMM is architected to host any number of tools that can conduct many more types of analyses on social media data.

We have mentioned Crimson Hexagon quite frequently thus far as Crimson is a web-based social media analytics platform with a rich datastore of social media data. Some issues with Crimson for an academic researcher are that it costs upwards of tens of thousands of dollars a year to use and that the data science models within the platform are largely "black box". One of the most valuable aspects of Crimson is their large datastore of social media data, such as their full index of Twitter data going back to 2010. The SMM could complement Crimson as we could build a connector within the SMM that connects to Crimson's datastore via Crimson's API. This would allow a researcher to use Crimson's vast datasets with the transparent data science models within the SMM.

Thus, we propose that the SMM helps to assist academic researchers who desire to conduct social media analytics without having a computational background. We believe that over time, the SMM will evolve to have a vast number of social media analytics tools, models, data, and researchers. To further explain the working dynamic between individual social media analytics tools and the SMM environment, we turn now to the architecture of the SMM.

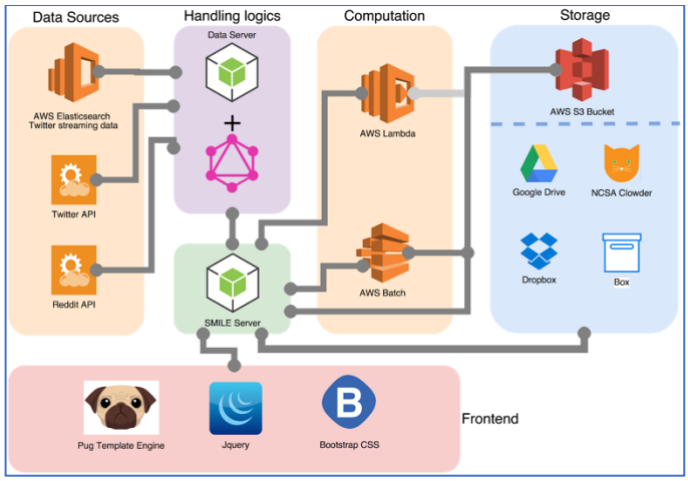


Fig. 3. The Social Media Macroscope Architecture.

## III. Architecture

The SMM is a web app that uses a virtual machine environment developed by HUBzero [11] at Purdue University. The HUBzero environment allows for the launching of tools in small customized virtual machine containers. This technique allows the tailoring of the environments to fit each tool including their software dependencies and keeps the tool sessions from interfering with one another. The HUBzero environment is hosted on an Amazon Web Services (AWS) EC2 virtual machine.

The first tool in the SMM is the Social Media Intelligence & Learning Environment (SMILE) which provides open source functions that collect social media data and analyze it. The tool currently provides access to Twitter and Reddit data and can perform text-preprocessing, sentiment analysis, network analysis and machine learning text classification. Future development of the SMM will add other social media collection and analysis tools and expand the capabilities of SMILE to include more functions and algorithms. Using the previous example of NodeXL, NodeXL could actually be containerized and then offered as a tool within the SMM. This could help to overcome the limitation of NodeXL only working on Windows platforms.

The AWS environment is extensively used by SMILE to enable easy scaling of compute resources and componentization of its analytical methods. As is pictured in Fig. 3, SMILE uses the EC2 machine only for requesting its functions. All of the computation and storage is accomplished by other AWS services. The main computation is done in AWS Lambda which is a serverless code service. Lambda houses the Twitter and Historical Reddit data collection scripts and versions of all 4 analysis methods. AWS Batch which is an on demand cluster environment houses versions of all of these scripts to be used for longer jobs as Lambda has a run limit of 5 minutes. The main Reddit search script is executed only in Batch as it has a long run time due to API rate limits. All of the storage of social data and results is done in AWS S3 which is a simple storage environment. Each Lambda or Batch script places its results in S3 and return a link to the S3 file to the main SMM EC2 instance.

## IV. SMILE Functions

SMILE offers numerous capabilities within the SMM, but we will only outline two for brevity sake, namely the capabilities of searching social media data and analyzing that data.

### A. *Searching Social Media Data*

One of the first processes that SMILE makes more accessible for SocRes is authenticating against public social media platforms' APIs (e.g., Twitter's API) and pulling data via search terms from that social media platform. Fig. 4. is a screenshot of how easy it is to authenticate against Twitter's API as compared to the processes outlined at https://developer.twitter.com/en/docs. Twitter users should be quite familiar to the authorization box as shown in Fig. 4.


Fig. 4. Twitter authorization screen within SMILE.

Fig. 5 shows the subsequent interface for searching Twitter data that is very similar to a Google type of search bar.


Fig. 5. Social media data search screen within SMILE.

Currently SMILE offers the ability to search Twitter and Reddit. These search capabilities are outlined next.

*1) Search Twitter:* SMILE offers two methods of searching Twitter for data via Twitter's API. Both methods are facilitated by using the GraphQL query language to interact with Twitter's REST API. Searching "Tweets" connects to Twitter's Search API endpoint to collect and return up to 180,000 posts from the last 7 days that match the keywords you provide. Searching "Twitter Users" connects to Twitter's User Search endpoint to collect and return up to 1,000 accounts that match the keywords you provide.

*2) Search Historical Twitter:* SMILE also offers the option to search our historical backlog of Twitter data. Our historical database includes tweets collected from Twitter's 1% streaming API since May 2018.

*3) Search Reddit:* SMILE offers three methods to search for Reddit data from the Reddit API. All three methods are facilitated by using the GraphQL query language.

- Search Reddit Posts connects to the Reddit API and returns the 1000 most recent posts containing the keyword given.
- Posts in Subreddit connects to the Reddit API to return the 100 most recent posts in a specific Subreddit.
- Comments in Subreddit connects to the Reddit API to return the 100 most recent comments in a Subreddit.

These processes can take a significant amount of time to complete for large numbers of posts due to limitations on API calls.

*4) Search Historical Reddit:* SMILE uses two methods to search for historical Reddit data. Both methods are facilitated by using the GraphQL query language to connect to Pushift.io's Reddit API.

- Historical Reddit Posts connects to the Pushshift.io endpoint for Reddit Posts to collect and return up to 10,000 Reddit posts who's titles match the keywords you provide.
- Historical Reddit Comments connects to the Pushshift.io endpoint for Reddit Comments to collect and return Reddit comments that match the keywords you provide. This can also be used once you completed a search for Reddit posts to return all of the comments to those posts.

Both of these processes take a significant amount of time to complete for large numbers of posts due to limitations on API calls.

## B. Analyzing Social Media Data

Fan and Gordon [5] list numerous data science methods in which researchers can analyze social media data, and the SMM provides a number of these methods in a format that is easy to use and transparent in methodology. We outline a few of the methods available within the SMM tool, SMILE.

*1) Sentiment Analysis:* SMILE provides sentiment analysis models with associated details as to how those sentiment analysis models were built. Fig. 6 provides a screenshot showing a pulldown menu in which a researcher could select which sentiment analysis model they would like to run against their social media data. In the case of Fig. 6, we have selected VADER, which is a sentiment analysis modeler built using Twitter data [7]. SMILE provides a citation for the paper behind VADER and it offers the ability to run the model without any need for programming. As an important note, many sentiment analysis models for social media data are built using one social media platform's data (e.g., Twitter data). When a researcher cannot see the details behind how the model was built, they may apply the model to data in which it was not trained on, therefore increasing the odds for erroneous results. Fig. 7 shows a screenshot of a portion of the results from SMILE's sentiment analysis using VADER, all of which can be downloaded to a CSV file.

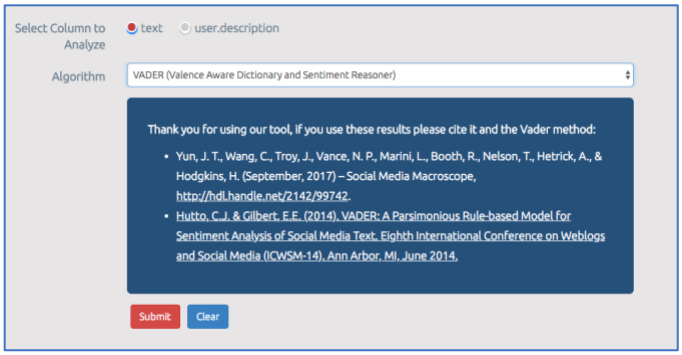


Fig. 6. Sentiment analysis menu within SMILE.

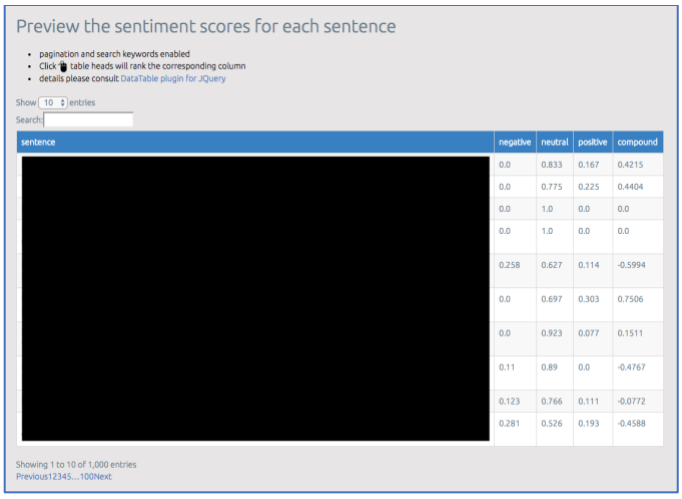


Fig. 7. Sentiment analysis results preview within SMILE (actual Tweets have been replaced by a black box to respect users' privacy).

Humphreys and Wang [12] recently suggested the use of VADER for consumer research in analyzing social text, but they did not provide any way to conduct this analysis outside of pointing to the conference paper for VADER [7]. Thus, we show how immediately helpful the SMM is for SocRes through a tool

like SMILE. As SMILE moves forward, future models for sentiment analysis will be added.

*2) Text Preprocessing:* SMILE provides text preprocessing capabilities which currently provides access to the Natural Language Toolkit (NLTK) Python Library [13]. This processing breaks post text down into phrases and words, removes stopwords, stems words, tags parts-of-speech and provides visual analysis of the most commonly used words and connections between word usage. As with sentiment analysis, future models for pre-processing will be built into SMILE.

*3) Text Classification:* SMILE currently provides supervised machine learning text classification powered by the Scikit-Learn Python Library [14]. This classification takes three separate tasks to create and test you prediction model.

- Step 1 splits your collected data into a training and testing set based on a ratio you provide and delivers the training set to you in CSV format. You will need to download that training set and note the training set identifying code that you are given. Then, you can label each post in the training set if the categories you would like to train your model to recognize.
- Step 2 requires you to upload your labeled training set and give your identifying code to match it to the saved testing set. This training set is used to create a machine learned model to identify posts that match the categories you trained.
- Step 3 Tests your model by using it to attempt to identify which posts in the training set match categories your model is trained to identify. You can then evaluate the successfulness of your model based on this test.

Training a machine learning model can take multiple attempts through this process before you get a successful model.

*4) Network Analysis:* SMILE also provides network analysis capabilities currently powered by the NetworkX Python Library [15] to help you analyze how content moved through the network of posters in your dataset. This tool identifies mention, reply and retweet interactions between users and can help you discover the influence of specific accounts on the network as a whole.

## V. Conclusion

As a community grows around the SMM, we envision additional tools, algorithms, and functionality to be added by members of the larger community. We envision that the SMM will bring us closer to a place where social researchers can focus primarily on the research question at hand, rather than all the technology details that surround the answering of that question.

## Acknowledgments

We would like to thank Technology Services at the University of Illinois for providing the necessary funding and support to make this project a reality. This would not have been possible without the support of Mark Henderson, John Towns, Tracy Smith, Tim Boerner, and Laura Herriott.

## References

[1] D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," *IEEE Intell. Syst.*, vol. 25, no. 6, pp. 13–16, 2010.

[2] "Global Social Media Analytics Market 2018 by Component, Mode of Deployment, End-User, Technology, New Innovation, Trends, and Forecasts to 2023," *Reuters*, 2018. [Online]. Available: https://www.reuters.com/brandfeatures/venture-capital/article?id=27472. [Accessed: 19-Jul-2018].

[3] A. Culotta and J. Cutler, "Mining Brand Perceptions from Twitter Social Networks," *Mark. Sci.*, p. mksc.2015.0968, 2016.

[4] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning," *Information*, vol. 9, no. 5, p. 127, 2018.

[5] W. Fan and M. D. Gordon, "Unveiling the Power of Social Media Analytics," *Commun. ACM*, vol. 12, no. JUNE 2014, pp. 1–26, 2013.

[6] "Leading Social Media Analytics Company | Homepage | Crimson Hexagon," 2017. [Online]. Available: https://www.crimsonhexagon.com/. [Accessed: 01-Jun-2017].

[7] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Eighth Int. AAAI Conf. Weblogs ...*, pp. 216–225, 2014.

[8] J. T. Yun and B. R. L. Duff, "Consumers as Data Profiles: What Companies See in the Social You," in *Social Media: A Reference Handbook*, 1st ed., K. S. Burns, Ed. Denver, Colorado: ABC-CLIO, LLC, 2017, pp. 155–161.

[9] "GitHub: cjhutto/vaderSentiment." [Online]. Available: https://github.com/cjhutto/vaderSentiment.

[10] M. A. Smith *et al.*, "Analyzing (social media) networks with NodeXL," in *Proceedings of the Fourth International Conference on Communities and Technologies*, 2009.

[11] M. McLennan and R. Kennell, "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering," *Comput. Sci. Eng.*, vol. 12, no. 2, pp. 48–53, Mar. 2010.

[12] A. Humphreys and R. J.-H. Wang, "Automated Text Analysis for Consumer Research," *J. Consum. Res.*, 2018.

[13] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," 2002.

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[15] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," 2008.