

# Mining Social Data with UCL's SocialSTORM Platform

R. Wood, I. Zheludev, and P. Treleaven

UK PhD Centre in Financial Computing, University College London, London, United Kingdom

**Abstract** - *SocialSTORM is a cloud-based 'central-hub' which facilitates the acquisition, storage and analysis of live data from social media feeds. Developed at University College London, the platform manages data from Twitter, Facebook, RSS sources and blogs, and is currently being extended to other feeds. This is for the purpose of harvesting information which ceases being publically-available shortly after creation. SocialSTORM includes facilities to run simulation models on the data allowing for the identification of changing trends, global sentiments and story propagation. Both historical and live data streams can be monitored. We are specifically interested in using these data for trading applications, although it has applicability to security monitoring, brand awareness and macroeconomic variable monitoring.*

**Keywords:** Social Science, Web Mining, Mining text and semi-structured data, Mining large scale data, Data mining software

## 1 Introduction

Social media is becoming a rich new area of research for scientific, sociological and commercial purposes. Acknowledging its huge value, it is likely that companies such as Google, Facebook and Twitter will increasingly restrict access to their social media data. With the rise of Computational Social Science, arguably what is required to support academic research is a public-domain social data scraping and analytics environment and high-performance computing facility.

To capitalize on the wealth of data currently available across the web, for the purposes of academic research, University College London (UCL) has built, and continues to develop, a comprehensive social media data engine that supports scraping and analysis of a wide range of social media data. This paper describes this social media Streaming, Online Repository and Analytics Manager (SocialSTORM) platform<sup>1</sup>. It can be seen as a multi-source customizable research-orientated counterpart to commercial social aggregation systems such as Datasift<sup>2</sup> and GNIP<sup>3</sup>. To our knowledge there are no similar social data acquisition and monitoring platforms tailored specifically to the research

community. The closest equivalent we have found is Wandora<sup>4</sup>, an open source information extraction, aggregation and data management system designed to run on a local machine. It is not directly suited to large-scale data mining, but via the creation of proprietary Java code, it can be used to monitor Twitter data.

## 2 Mining large scale data

### 2.1 Web mining for social data

Currently, large quantities of public data from sources such as Twitter and Facebook can be acquired free of charge. Data are typically accessed by querying an Application Programming Interface (API) for each of these respective social media providers; and this may be used to tailor results according to a desired dataset via proprietary code. Twitter for example, allows developers to track up to 400 specified keywords for which to filter publicly available updates before streaming to the developer in near real-time. It is also possible to filter Twitter data by user ID or location, achievable with a simple HTTP POST request. Obtaining a 'random sample' of data from Twitter is even easier; the following HTTP GET request returns a live stream roughly 1% of all public status updates as a JSON array:

<https://stream.twitter.com/1/statuses/sample.json>

Furthermore, elevated access to a random sample of approximately 10% of all global Tweets is straightforward to obtain for academic research purposes. Once these Twitter data have been published and streamed through its API, the data ceases to be accessible from Twitter. This highlights the need for continuous communication with Twitter, and suitable technologies for storage of the data to allow aggregation of a substantial dataset over time (discussed later).

Facebook also offers an API through which publicly available data are accessible; though not in real-time. It is also less common for Facebook users to make their updates publicly visible – this is in contrast to Twitter's policy where Tweets are automatically in the public domain by default (with an opt-out option). However, given Facebook's user-base of c. 800 million people, it is still reasonable to expect large volumes of data to be available for retrieval and

---

<sup>1</sup> [www.socialstorm.eu](http://www.socialstorm.eu)

<sup>2</sup> [www.datasift.com](http://www.datasift.com)

<sup>3</sup> [www.gnip.com](http://www.gnip.com)

---

<sup>4</sup> [www.wandora.org](http://www.wandora.org)

analysis. Facebook’s Graph API can be used to search for status updates containing particular keywords specified by the developer; these results go back as far as 70 days from the request date. The following HTTP GET request can be used to access these data, returning a JSON array of data relating to all public posts containing the term ‘Apple’ used here as an example:

<https://graph.facebook.com/search?q=Apple&type=post>

Through the same API it is also possible to retrieve public data relating to Facebook *Pages, Events, Users, Groups, Places* or *Checkins* by modifying the ‘type’ parameter in the above URL accordingly.

A ‘random sample’ of all public updates from Facebook can also be harvested, by using the query search function to look for a collection of a language’s most commonly used words such as: ‘to’, ‘be’, ‘and’, ‘of’ if working in English.

## 2.2 Mining text and semi-structured data

Before analyzing social media data from Facebook and Twitter, one must extract the relevant data fields from their raw JSON format. Fig. 1 provides an example of the information fields returned for each Tweet retrieved via the filtering method of Twitter’s Streaming API (data has been removed to protect privacy).

```
{
  "text": "",
  "entities": {},
  "contributors": ,
  "place": ,
  "id_str": " ",
  "coordinates": ,
  "source": " ",
  "retweet_count": ,
  "in_reply_to_user_id": ,
  "in_reply_to_status_id": ,
  "favorited": ,
  "geo": ,
  "in_reply_to_screen_name": " ",
  "truncated": ,
  "in_reply_to_status_id_str": " ",
  "user": {},
  "retweeted": ,
  "id": ,
  "in_reply_to_user_id_str": " ",
  "created_at": " "
```

Fig. 1. Example response from Twitter’s Streaming API

Twitter offers many forms of metadata which may also provide a source for analysis, as well as the Tweet (“text”) itself. Examples include location tags and the number of times the message was “retweeted” (shared by other users, thus increasing the audience). These may consist of integers, strings, or a combination of both. In order to extract these data, one needs to parse the raw JSON structure and store each desired string or integer as a separate variable. Conveniently, one may use the very same method to structure data retrieved from Facebook (which is also in JSON by

default). The data may then, for example, be stored within separate columns of a database, or as a text file with a specified delimiter. It is inadvisable to store this kind of data as a text file in CSV (comma-separated variable) format since status updates themselves frequently contain one or more commas which can make subsequent analysis tricky.

Text data may be analyzed in a number of ways, using techniques from Natural Language Processing and Information Retrieval. An obvious direction to take when analyzing text-based social media data is to conduct sentiment analysis in order to quantify message strings, to enable mathematical models to be employed for analysis.

## 3 SocialSTORM

### 3.1 Overview

University College London, assisted by Microsoft, has built a cloud-based computational finance environment (ATRADE<sup>5</sup>) that supports real and virtual trading; with terabytes of financial data to support research into algorithmic trading and risk. Given the rise of interest in using social data (e.g. Twitter updates) for trading and risk management, UCL has now built SocialSTORM, a complementary social media engine that supports scraping and analysis of a wide range of social media data.

As discussed, SocialSTORM is a cloud-based platform which facilitates the acquisition of text-based data from online sources such as Twitter, Facebook, respected blogs, RSS media and ‘official’ news; a ‘central-hub’ for social media analytics. The system includes facilities to upload and run Java-coded simulation models to analyze the aggregated data; which may comprise UCL’s social data and/or users’ own proprietary data. There is also connectivity to the ATRADE platform which provides further quantitative finance and economic data.

The platform consists of infrastructure and tools to facilitate data acquisition, database connectivity, and various levels of access and administration along with data repositories for long and short-term data storage. The platform is able to operate in two simulation modes: an ‘historical’ mode which utilizes data already stored at the time of running the desired simulation (ideal for data-mining and back-testing), and a ‘live’ mode which operates on a near real-time stream of data which is continually monitored from the sources throughout the simulation (ideal for analyzing financial markets and developing algorithmic trading strategies).

<sup>5</sup> <http://vtp.cs.ucl.ac.uk/atrade>

In short, SocialSTORM allows for the execution of user-defined simulation models for the analysis of historical and real-time data feeds that provide a plentiful supply of public opinions derived from online-community data.

### 3.2 Infrastructure architecture

The SocialSTORM platform resides in a distributed computing environment currently consisting of 9 nodes each with the following specification: 15,000rpm 600GB hard drive, 32GB RAM and one 3.2GHz quad-core Intel Xeon e3-1200 processor. The nodes are interlinked by 10GbE (10 Gigabit Ethernet) connections and the entire system is backed-up daily onto tape storage for up to 3 months. SocialSTORM's current storage capacity is 5.4TB with 288GB of available RAM. SocialSTORM is fully scalable – additional nodes can be added to increase system storage and performance on an as-needed basis.

This particular hardware setup has been chosen for the purposes of migrating SocialSTORM to Apache Hadoop, a software library and framework that allows for the distributed processing of large data sets<sup>6</sup>; which is something we are currently working towards.

### 3.3 System architecture

SocialSTORM inherits its architectural design from UCL's ATRADE system, which allows easy integration between the two systems. The following is an outline of the key components of the SocialSTORM system.

**Connectivity Engines** – Various connectivity modules communicate with the external data sources, including Twitter & Facebook's APIs, financial blogs and various RSS news feeds; and are being continually expanded to incorporate new social media sources. Data are fed into SocialSTORM in real-time and include a random sample of all public updates from Twitter, as well as filtered data streams selected from a rich dictionary of stock symbols, currencies and other economic keywords; providing gigabytes of text-based data every day.

**Messaging Bus** – This serves as the internal communication layer which accepts the incoming data streams (messages) from the various connectivity engines, parses these (from either JSON or XML format) and writes the various data to the appropriate tables of the main database.

**Data Warehouse** – This is home to terabytes of text-based entries which are accompanied by various types of metadata to expand the potential avenues of research. Entries are organized by source and accurately time-stamped with the time of publication, as well as being tagged with topics for easy retrieval by simulation models.

**Simulation Manager** – This terminal provides the external API for clients to interact with the data for the purposes of analysis, including a web-based GUI via which users can select various filters to apply to the datasets before uploading a Java-coded simulation model to perform the desired analysis on the data. The Simulation Manager facilitates all client-access to the data warehouse, and also allows users to upload their own datasets for aggregation with UCL's social data for a particular simulation. There is also the option to switch between historical mode (which analyses data existing at the time the simulation is started) or live mode (which 'listens' to incoming data streams and performs analysis in real-time).

In summary, the aims of SocialSTORM include acquisition and access to terabytes of social data from a variety of sources, as well as a cloud-based simulation environment for historical data-mining and real-time monitoring of global news and opinions taken from the world's most popular social networking sites. There is also connectivity to UCL's ATRADE algorithmic trading system and support for aggregation of clients' proprietary data. UCL's cloud-based platform removes the need to transfer large amounts of data across servers and also eliminates dependency on the processing power of clients' local machines; leading to increased performance in working with 'Big' datasets.

### 3.4 Data storage

SocialSTORM queries and monitors social media APIs in real-time, reading updates as they are streamed and writing these directly to its database. The latency between a message being published to Twitter (as an example) and subsequently being stored in our database is less than 1 second; even when using batch inserts to increase efficiency. Typically, the system writes c. 4,000 entries to the database every second.

From Twitter the current system retrieves c. 20 million messages per day as a 'random sample' of all public updates, plus c. 1-2 million messages daily containing hundreds of specific financial and economic keywords selected by the platform's development team. From Facebook, a proprietary method of retrieving a random sample of all public updates is used which returns c. 2 million updates per day. This is supplemented by searching for updates containing the same keywords used to filter updates from Twitter; giving over 500,000 additional daily updates from Facebook. The SocialSTORM team has selected 15 finance-related blogs to monitor, as well as a number of official news services which, together contribute over 1,000 daily entries to the database.

The current data sources result in approximately 5GB of data per day, which is likely to continue to increase over time barring any restriction to public data by Social Media companies; UCL has the facilities to cope with an increased dataflow. The current SocialSTORM servers allow storage of

---

<sup>6</sup> <http://hadoop.apache.org/>

multiple terabytes of data; but may be scaled-up to petabytes if required.

### 3.5 Simulations

User-privacy is taken very seriously by the platform’s development team. Although the data retrieved from the web is in the public domain, it remains property of the data provider and is therefore not redistributable in accordance with Content License Agreements. To enable analysis of social media data by third parties, SocialSTORM includes a black-box research environment, accessible via a graphical web interface as shown in Fig. 2. Here, subscribed users may upload their own java-coded simulation models which will analyze the data stored by SocialSTORM and return post-analytical results to the caller according to the model used.

Models to be uploaded to SocialSTORM are **.jar** files, which also include any packages on which the code is dependent. The simulation environment then looks for a particular method, similar to **Main()**, which defines the appropriate parameters to interface with SocialSTORM’s API. Instructions on how to ensure that models are compliant with the platform are detailed in the SocialSTORM user manual.

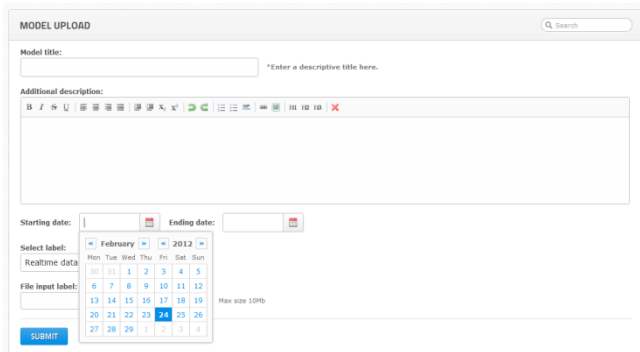


Fig. 2. Model-upload form for SocialSTORM simulations

Before running a model the user can opt to perform the analysis in ‘live’ mode, which connects directly to the platform’s real-time messaging system to stream live updates to the model, or ‘historical’ mode which retrieves data already stored in the database. The user may choose to pause or stop a simulation at any time, and a ‘live’ simulation is complete when a certain breakpoint in the code is reached or until the user manually stops the simulation. Once a simulation is complete, users can plot results in various ways using the SocialSTORM GUI (an example of which is shown in Fig. 3), export results to Microsoft Excel, or use an output API to retrieve the results programmatically for further analysis. Data exported to Microsoft Excel can be linked to constantly update in a spreadsheet’s cells.

## 4 Applications

The applications of social media data in academia and commerce are growing rapidly. As Computational Social

Science expands, platforms such as SocialSTORM should provide useful research tools. To demonstrate the wider appeal of aggregated social media data, we present the following specific examples.

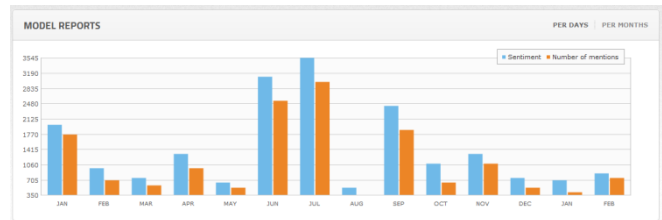


Fig. 3. Example of a bar chart plot of simulation results in SocialSTORM’s web interface

### 4.1 Social science

The analysis of narratives of professional journalistic articles for the demonstration of ‘Phantastic Objects’<sup>7</sup> and their effect is of growing importance to the understanding of financial and economic bubbles [2]. As an example of the application of social data to the social sciences, it is a logical next step to consider Phantastic Objects in every-day social media data. Such analyses could lead to new findings in the realm of ‘meme’ propagation, with a better understanding of why sentiment bubbles occur.

To achieve this, a manageable database of social media data is required, and with its data analytics capabilities, SocialSTORM is able to provide this stepping stone.

### 4.2 Business intelligence

A growing number of services provide business intelligence to firms seeking to monitor the sentiment around their company’s name on the internet. Furthermore, sentiments around products influence future product design, and real-time analysis of the world’s ‘mood’ on a brand or company dictates the success of digital marketing endeavors.

However, a particular aspect of digital business intelligence is currently in its infancy. It is becoming increasingly apparent that the instant global sentiment around a firm or its products can be used as a predictor of that firm’s future performance. Whilst previously establishing such relationships was only possible after the time-consuming analysis of professionally-written publications on a company’s performance or its products, it is now clear that public ‘groupthink’ opinion is just as relevant. The analysis of social media data in a near-instant capacity is thus the methodology needed, and SocialSTORM can facilitate this process.

<sup>7</sup> Stemming from the word ‘phantasy’, this term is used in the sense of meaning an imaginary scene in which the inventor of the phantasy is a protagonist in the process of having his or her latent (unconscious) wishes fulfilled [1].

### 4.3 Advanced prediction modeling

The prediction and/or estimation of macroeconomic variables such as consumer confidence, unemployment, and inflation are of great importance to both policy makers and investors. Current methodologies of observing such variables are highly limited. Not only are they survey-based, meaning that the data-accumulation and analysis process is extremely time and cost intensive, but the results are often cited in literature as being simply inaccurate. Much value therefore lies in the timely prediction of such macroeconomic variables, and it has already been shown that the analysis of sentiment derived from search engine data can be used as a predictor for financial and economic variables [3].

Thus, it is of interest to explore the effect of sentiment variation of social media data on macroeconomic variables such as those mentioned above. With SocialSTORM's capacity to monitor both historic and live data via the use of custom-written models, such analyses may now be feasible.

## 5 Future work

SocialSTORM is in constant development, and the following additional features are in the process of being implemented:

**Sentiment classification** – A series of in-built machine-learning packages are being developed which will allow for the sentiment analysis of the text stored in SocialSTORM's database. This analysis will allow users to quantitatively rank social media text based on emotions such as anger, anxiety, happiness and sadness.

**Esper** – This is a complex event processing package, which allows for the high-speed processing of large volumes of events in real-time<sup>8</sup>. The integration of this package will implement SocialSTORM's sentiment classification algorithms in real-time, before messages are stored in the database; and will also improve the system's overall performance when simulation models are operating on live data.

**Advanced Visualization Suite** – The current Web interface is being improved and expanded to allow for the customization of the visualization of the output data produced by the simulation models.

**Integration with ATRADE** – Upon completion, this functionality will allow SocialSTORM's users to run models that can simultaneously evaluate financial data from the ATRADE platform, as well as its native social media data.

## 6 Conclusion

UCL's SocialSTORM platform is a data mining and analytics engine that can provide access and customized monitoring capabilities for aggregated social media. Being a non-commercial product, the platform offers researchers a facility to monitor and evaluate a rich and yet often fleeing data source. The platform complements existing commercial products which offer similar capabilities, but are not primarily targeted at the wider academic community. SocialSTORM's customizable nature also allows for integration with local software to support research in Computational Social Science.

## 7 References

- [1] J. Laplanche, J. Pontalis The language of psychoanalysis. Nicholson-Smith D, translator. New York, NY, London: Norton and Hogarth, 1973.
- [2] D. Tuckett, R. Taffler, "Phantastic objects and the financial market's sense of reality: a psychoanalytic contribution to the understanding of stock market instability", *Int. J. Psychoanal.*, vol. 89, 389–412, Jun 2008.
- [3] H. Mao, S. Counts, J. Bollen, "Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data (Periodical style–Submitted for publication)", eprint arXiv:1112.1051, submitted for publication, Dec. 2011.

---

<sup>8</sup> <http://esper.codehaus.org/>