Wildfire: A Twitter Social Sensing Platform for Layperson

Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang
Foram Patel, Josue Caraballo, Patrick Hennecke, Chengkai Li
University of Texas at Arlington, USA
{zeyu.zhang,zhengyuan.zhu,haiqi.zhang,fxp3176,josue.caraballo,pwh8492}@mavs.uta.edu,cli@uta.edu

ABSTRACT

We present Wildfire, an innovative social sensing platform designed for laypersons. The goal is to support users in conducting social sensing tasks using Twitter data without programming and data analytics skills. Existing open-source and commercial social sensing software only supports data collection using simple keyword-based search. On the contrary, Wildfire employs a heuristic graph exploration algorithm to selectively expand the collected tweet-account graph in order to further retrieve more task-relevant tweets and accounts. This approach allows collecting data to support complex social sensing tasks that cannot be met by simple keyword search. In addition, Wildfire provides a range of analytic tools, such as text classification, topic generation and entity recognition, which can be crucial for tasks such as trend analysis. The platform also provides a web-based user interface for creating and monitoring tasks, exploring collected data, and performing analytics.

PVLDB Reference Format:
Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang
Foram Patel, Josue Caraballo, Patrick Hennecke, Chengkai Li. Wildfire: A Twitter Social Sensing Platform for Layperson. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

1 INTRODUCTION

Social media has become an integral part of our lives, reflecting people's opinions, behaviors, and interactions. It offers a means to observe and interpret phenomena and discover insights about our society. The practice of doing that, called *social sensing* [7], can be applied by health offices to surveil epidemic outbreaks, by social scientists to understand crowd behavior, and by governments to detect crimes. Social sensing offers opportunities beyond traditional methods. It can scale up to millions of users within just a few days, in contrast to traditional surveys, and it can capture real-time changes in public opinion, which sets it apart from traditional polls. Moreover, social sensing offers unique insights from data that only exist in social media, e.g., "likes". By enabling sophisticated, large-scale analysis of social media data, technological advancements in deep learning and natural language processing (NLP) have greatly expanded the benefits of social sensing.

Those who need to conduct social sensing in the real world are non-technical users, but social sensing requires complex skills in coding, data modeling, data processing, and analytics. Social media data can be abstracted as a graph with posts, users, and their relationships. It is non-trivial to retrieve the data from platforms (e.g., through various Twitter APIs, [1] using search and query conditions), to parse the data which could be in formats such as JSON and may contain both texts and images, to store them in accordance with well-designed schema, and to serve the stored data based on search and filter conditions. All these are particularly challenging given the large volume and velocity of social media. Given practical constraints such as the access rate limits in Twitter APIs, one often needs to devise sophisticated algorithmic approaches to prioritizing which part of the graph to retrieve before others. Once the data are in place, various types of content analysis (e.g., text classification, sentiment analysis, stance detection), network analysis (e.g., ranking users based on importance), and data visualization are performed. These analytics may need to be repeated periodically given the constantly updated data.

This paper introduces Wildfire (demo available at https://idir.uta.edu/wildfire), a novel social sensing platform focusing on Twitter data. It enables anyone with an academic Twitter token [2] to conduct social sensing tasks on large volumes of data, without writing a single line of code.

- For *data collection*, Wildfire provides two methods. One is *seed collection*, which uses Twitter's Stream API to collect random tweets or Search API to collect tweets that match user-provided keyword search or query. This method suffices for simple tasks such as collecting tweets containing a target hashtag. Another method is *expansion collection*, which brings in additional Twitter accounts and tweets on top of collected data. This entails using Twitter's Timeline API to collect an account's past tweets and using a customized ranking function to select the most relevant Twitter account among collected accounts, followed by obtaining its followees (i.e., the accounts it follows) using Twitter's Following API. This method supports complex tasks such as user behavior profiling, where it helps to discover similar accounts through the following relationships on Twitter.
- For *data analytics*, Wildfire offers a range of tools such as text classification, topic generation, and entity recognition. These analytics tools allow users to generate classification scores, extract topic labels, and identify entities from collected tweets. The results of such analytics benefit social sensing tasks. For instance, in trend analysis, topic generation is useful in uncovering popular topics among a certain population.
- Wildfire uses a single graph to store all Twitter accounts and tweets collected for multiple tasks of the same user. Meanwhile collected data are annotated so that which part of the graph

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

[1] https://developer.twitter.com/docs/twitter-api

[2] https://developer.twitter.com/products/twitter-api/academic-research

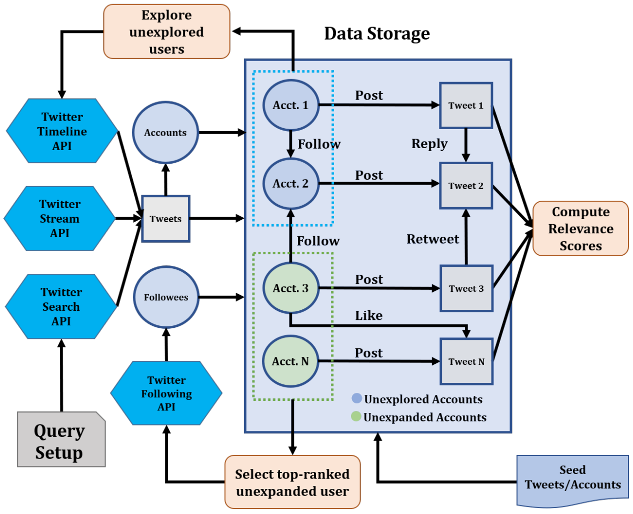


Figure 1: Data collection component architecture

belongs to each specific task becomes clear. Using this approach Wildfire avoids redundant collection and storage of data.

The distinguishing features of Wildfire are its expansion collection, which allows for extensive collection of task-relevant accounts and tweets, and analytics tools, which enable users to gain insights and make decisions based on patterns and trends that emerge from collected data. Open-source tweet collection tools such as [1, 4, 6] lack expansion collection capabilities and only provide basic data statistics, such as the frequency of hashtags and mentions. While commercial social sensing tools (e.g., brandwatch.com, digimind.com, hootsuite.com, and synthesio.com) offer comprehensive data analytics, they do not expand tweet collections based on the following relationships. In contrast, Wildfire provides a highly adaptable expansion collection feature since users can customize the ranking function by selecting and prioritizing specific features, making it suitable for a wide range of tasks.

2 DATA COLLECTION BACKEND

Figure 1 shows the architecture of Wildfire's data collection component. The collected tweets and Twitter accounts conceptually form a bipartite graph. The edges in the graph correspond to relationships between accounts and tweets, including Twitter accounts posting and liking tweets, accounts following each other, and tweets retweeting, quoting, and replying-to other tweets. The graph is stored in a MySQL database.

To use Wildfire, a user is required to sign up and provide an academic Twitter token. Once logged in, the user can create multiple social sensing tasks. The system uses the user-provided Twitter token to collect data for their tasks. Wildfire accommodates concurrent execution of multiple tasks from the same user. Instead of populating a separate graph for each task, Wildfire stores data across all tasks of the user in a single graph. This avoids wasteful, redundant data retrieval and storage. Particularly, actions such as retrieving timeline tweets and follower/followee lists for a specific account will only be performed once, even if the account is included in the graphs for multiple tasks. This novel and unique design allows Wildfire to share data across a user's tasks, resulting in reduced usage of the user's token and shortened task completion time.

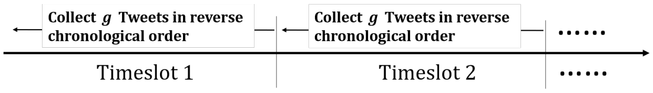

Collect g Tweets in reverse chronological order
Collect g Tweets in reverse chronological order
......
Timeslot 1
Timeslot 2
......

Figure 2: Timeslot and granularity

Given the aforementioned system design, the ensuing discussion in this section focuses on how Wildfire collects data from Twitter for one task. For that, the system populates the graph with two methods that use Twitter APIs, as follows.

1) The *seed collection* method uses the Search API to obtain tweets, based on user-specified keywords or Twitter queries,[3] and further obtains the corresponding Twitter accounts. Alternatively, the user can choose to use the Stream API to randomly retrieve tweets.

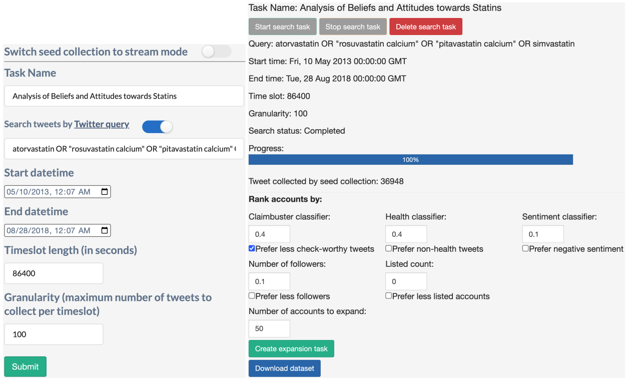
Since Twitter limits the number of historical tweets that can be retrieved each month and returns tweets in reverse chronological order, Wildfire requires a user to specify a few parameters when creating a task, in order to avoid rapid exhaustion of the user's monthly quota, meanwhile ensuring evenly distributed tweets across a task's search period. The parameters include *Start datetime* (s), *End datetime* (e), *Timeslot length* (t), and *Granularity* (g). More specifically, Wildfire splits the search period $[s, e]$ into $\frac{e-s}{t}$ timeslots. For each timeslot, it collects at most g tweets. Figure 2 illustrates the idea. Note that *Start datetime* and *End datetime* can be in the future. Before collecting tweets for each timeslot, Wildfire checks whether the end of the timeslot is past. If not, it sleeps until the end of the timeslot and then uses the Search API to retrieve past tweets in reverse chronological order.

2) The *expansion collection* method uses the Timeline API to *explore* a Twitter account by retrieving up to 3,200 most recent tweets in its timeline and the Following API to *expand* the account by retrieving its following list (i.e., the list of accounts that this account follows). We refer to the accounts for which we have not collected timeline tweets as *unexplored accounts* and the accounts for which we have collected timeline tweets but not following lists as *unexpanded accounts*.

The expansion collection method operates with two concurrent, continuous processes. One process always randomly chooses an unexplored account to explore, i.e., collecting its timeline tweets. Once an account is explored, it becomes an unexpanded account. The other process employs a customized *ranking function* to select the most relevant unexpanded account to expand in hopes of finding additional accounts and tweets that are relevant to the given social sensing task. The rationale behind the ranking function is that the followees of a relevant Twitter account may tend to be relevant as well since they may share common interests and beliefs and they may participate in joint or similar conversations.

Wildfire enables users to use multiple weighted property scores to construct their ranking functions. Those property scores can be some simple account attributes such as the followers count and listed count (number of Twitter lists the account is in), or some

[3] https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query



Switch seed collection to stream mode

Task Name

Analysis of Beliefs and Attitudes towards Statins

Search tweets by Twitter query

atorvastatin OR "rosuvastatin calcium" OR "pitavastatin calcium" (

Start datetime

05/10/2013, 12:07 AM

End datetime

08/28/2018, 12:07 AM

Timeslot length (in seconds)

86400

Granularity (maximum number of tweets to collect per timeslot)

100

Task Name: Analysis of Beliefs and Attitudes towards Statins

Query: atorvastatin OR "rosuvastatin calcium" OR "pitavastatin calcium" OR simvastatin

Start time: Fri, 10 May 2013 00:00:00 GMT

End time: Tue, 28 Aug 2018 00:00:00 GMT

Time slot: 86400

Granularity: 100

Search status: Completed

Progress: 100%

Tweet collected by seed collection: 36948

Rank accounts by:

Claimbuster classifier: 0.4

☑Prefer less check-worthy tweets

Health classifier: 0.4

□Prefer non-health tweets

Sentiment classifier: 0.1

□Prefer negative sentiment

Number of followers: 0.1

□Prefer less followers

Listed count: 0

□Prefer less listed accounts

Number of accounts to expand: 50

Figure 3: (a) Task creation (b) Task monitoring and expansion

deeper account features such as sentiment and political spectrum measurements, which can be computed by sophisticated classifiers. Currently, we provide several task-specific classifiers: *ClaimBuster* [3] - a deep learning model to detect tweets containing check-worthy factual claims; *Health tweets detector* [4] - a self-developed deep-learning model to identify tweets related to health; *Sentiment Analyzer* [5] - a multilingual model to discover the sentiment of tweets. Each classifier can compute a relevance score for a tweet to determine its relevance under this classifier, while the corresponding property score for the Twitter account is the average relevance score of all its timeline tweets. By assigning weights to those property scores, the user can construct its own ranking functions. Note that each weight should fall between 0 and 1, and users can use lower property scores if preferred. When computing the final ranking grade, normalization will be applied to fit property scores into [0,1] to uniform scales, then a weighted average of all property scores will be computed to obtain the final ranking grade.

Once the user has selected the property scores and their corresponding weights for the ranking function, as well as specifying the desired total number of accounts to expand, the process of creating the expansion collection is complete. This expansion collection can be run concurrently with the seed collection, thereby accelerating the data collection process.

3 USER INTERFACE

The user interface comprises three main components, namely, a *task creation* page (Figure 3(a)), a *task monitoring and expansion* page (Figure 3(b)), and a *data analytics* page (Figure 4). The task creation page prompts the user for the input parameters to initiate a task. The task monitoring and expansion page displays the status and progress of the user's tasks and allows the user to control each task and configure its expansion. The data analytics page of each task allows the user to explore the collected tweets using multiple filters and examine corresponding analytics results.

Task creation (Figure 3(a)) The user can toggle between search mode (i.e., using Search API) and stream mode (i.e., using Stream API) for seed collection. For the steam mode, the user only needs to specify the number of tweets to collect. For the search mode, the user can further toggle between simple (i.e., keywords) and advanced search (i.e., query) modes. The keywords mode allows the user to enter multiple keywords separated by commas, in which a keyword may consist of multiple tokens. The query mode allows the user to use a Twitter query. [3] The user can also specify *Start datetime*, *End datetime*, *Timeslot length*, and *Granularity*, as discussed in Section 2.

Task monitoring and expansion (Figure 3(b)) For each task, this page displays its configuration, status (e.g., active, completed, stopped, and error), and collection progress. The user is given options to start, stop, resume, or delete the task. In creating an expansion on the task, the user will choose the weight and preference associated with each ranking factor (e.g., classification scores, follower count, list count) and enter the number of Twitter accounts to expand. The user is also given the start, stop, resume, and delete options on the expansion. The "Download" button at the bottom of each task leads to a page where the user can choose what to include in the downloaded dataset, e.g., whether to include the profile of each Twitter account.

Data analytics (Figure 4) In order to filter and analyze the collected data, Wildfire provides a search engine with various filtering options. Users can filter tweets based on the task, keywords, hashtags, from/mentioning specific accounts, date range, and classification score range for more accurate categorization. Once the desired filters are applied, the corresponding result tweets are displayed and paginated. The analytic tools include topic generation [6] for generating thematic tags of the tweets, entity recognition [7] for extract entities from tweets, and the aforementioned three classifiers. The first two are powered by models from Hugging Face and hosted on the Wildfire backend. There are two ways for the user to obtain analytic results. The first method involves clicking one of the three result buttons: classifier results, topic generation results, and entity recognition results. This displays the aggregated results derived from all the filtered tweets, including three classification score distributions presented as a pie chart (the blue frame in Figure 4), a bar chart displaying the top 10 most frequent tags representing the topics discussed among the tweets, and another bar chart showing the top 10 most frequent entities recognized from the tweet text. The second way is to click on a single tweet displayed at the bottom of the page, which opens a pop-up window (orange frame in Figure 4) showing the results of all five tools applied to the tweet. The output of the three classifiers is presented as a score between 0 and 1. The topic generation tool generates a list of topic words derived from the tweet text, even if these words do not necessarily appear in the original tweet. The results of entity recognition are provided as pairs of entities and their corresponding entity types.

4 DEMONSTRATION SCENARIO

We present two use cases that illustrate how individuals without specialized research skills can utilize the Wildfire to effectively replicate and even extend existing research.

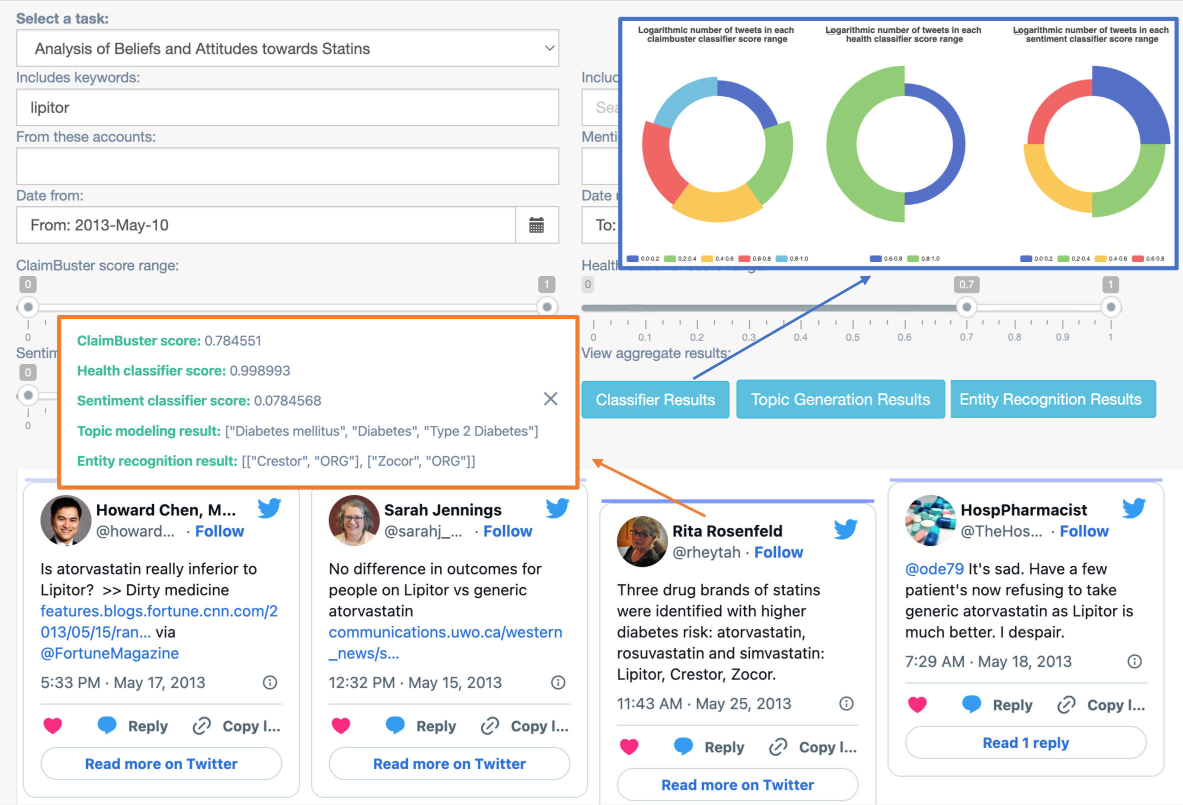
Use Case 1: Analysis of Beliefs and Attitudes towards Statins. The authors of [2] examined people's attitudes and beliefs toward statins, a class of medications used to reduce cholesterol. The study involved a collection of 12,649 tweets, which were manually categorized

[4] https://idir.uta.edu/health_tweet_classifier

[5] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[6] https://huggingface.co/fabiochiu/t5-base-tag-generation

[7] https://huggingface.co/autoevaluate/entity-extraction



Select a task: Analysis of Beliefs and Attitudes towards Statins

Includes keywords: lipitor

From these accounts:

Date from: From: 2013-May-10

ClaimBuster score range: 0 – 1

Health classifier score range: 0.7 – 1

View aggregate results:

Classifier Results | Topic Generation Results | Entity Recognition Results

ClaimBuster score: 0.784551
Health classifier score: 0.998993
Sentiment classifier score: 0.0784568
Topic modeling result: ["Diabetes mellitus", "Diabetes", "Type 2 Diabetes"]
Entity recognition result: [["Crestor", "ORG"], ["Zocor", "ORG"]]

Howard Chen, M... @howard... · Follow
Is atorvastatin really inferior to Lipitor? >> Dirty medicine features.blogs.fortune.cnn.com/2013/05/15/ran... via @FortuneMagazine
5:33 PM · May 17, 2013

Sarah Jennings @sarahj_... · Follow
No difference in outcomes for people on Lipitor vs generic atorvastatin communications.uwo.ca/western_news/s...
12:32 PM · May 15, 2013

Rita Rosenfeld @rheytah · Follow
Three drug brands of statins were identified with higher diabetes risk: atorvastatin, rosuvastatin and simvastatin: Lipitor, Crestor, Zocor.
11:43 AM · May 25, 2013

HospPharmacist @TheHos... · Follow
@ode79 It's sad. Have a few patient's now refusing to take generic atorvastatin as Lipitor is much better. I despair.
7:29 AM · May 18, 2013

Logarithmic number of tweets in each claimbuster classifier score range
0.0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
Logarithmic number of tweets in each health classifier score range
0.6-0.8
0.8-1.0
Logarithmic number of tweets in each sentiment classifier score range
0.0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0
1
0.7
0.1
0.2
0.3
0.4
0.5
0.6
0.8
0.9
Reply
Copy l...
Read more on Twitter
Read 1 reply

Figure 4: Data analytics page

into ten topics to reveal the public reaction about statins on social media. Wildfire can potentially help facilitate the data collection and analysis. We created a keywords search collection, as shown in Figure 3(a), and retrieved 36,948 tweets related to statins. To look into people's attitudes towards Lipitor, we first filtered collected tweets with the keyword "lipitor" and a high health score range [0.7,1] on the data analytics page (Figure 4), resulting in 2,457 tweets. The distribution of sentiment classification scores shows that 2,205 tweets, accounting for approximately 90% of the filtered tweets, contain negative sentiment with scores lower than 0.2.

Use Case 2: How People Make Suicide Pacts on Twitter. To investigate whether individuals utilize Twitter to search for like-minded persons with suicidal ideation and form suicide pacts, [5] collected tweets that contain "suicide pacts" in Korean from Oct 16th to Nov 30th, 2017. We created a search collection targeting the same keywords and incorporated an expansion that applies a negative sentiment classifier in the ranking function. The expansion brought us more relevant tweets and accounts by expanding the seed accounts collected by the search collection. The search collection retrieved 10,953 tweets and its expansion collected 2,051 tweets from Feb 1st, 2023 to Mar 24th, 2023. Among all of them, we found 11,312 tweets that contained the keyword "suicide pacts," which is significantly more than what was retrieved by the search collection alone. This indicates that the expansion is a valuable tool for gathering relevant data in this particular case, perhaps because individuals seeking to form suicide pacts are more likely to follow one another.

5 FUTURE WORK

We plan to implement several new functions to enhance the platform's utility: 1) Besides creating seed collections, users can upload a file containing seed data. 2) In addition to the pre-installed classifiers that we offer, users can also upload their own models as scorers. 3) The data analytics page will include more filtering options, such as excluding specific keywords and accounts. Moreover, more analytic tools, such as stance detection and bot detection, will be added.

REFERENCES

[1] Erik Borra and Bernhard Rieder. 2014. Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib journal of information management* 66, 3 (2014), 262–278.

[2] Su Golder, Karen O'Connor, Sean Hennessy, Robert Gross, and Graciela Gonzalez-Hernandez. 2020. Assessment of beliefs and attitudes about statins posted on Twitter: a qualitative study. *JAMA network open* 3, 6 (2020), e208953–e208953.

[3] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD*. 1803–1812.

[4] Michael W Kearney. 2019. rtweet: Collecting and analyzing Twitter data. *Journal of open source software* 4, 42 (2019), 1829.

[5] Sang Yup Lee and Yeji Kwon. 2018. Twitter as a place where people meet to make suicide pacts. *Public Health* 159 (2018), 21–26.

[6] George Washington University Libraries. 2016. Social Feed Manager. (2016).

[7] Dong Wang, Boleslaw K. Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. 2019. The Age of Social Sensing. *Computer* 52, 1 (2019), 36–45.