Wildfire: A Twitter Social Sensing Platform for Layperson

Zeyu Zhang
University of Texas at Arlington
zeyu.zhang@mavs.uta.edu

Zhengyuan Zhu
University of Texas at Arlington
zhengyuan.zhu@mavs.uta.edu

Haiqi Zhang
University of Texas at Arlington
haiqi.zhang@mavs.uta.edu

Foram Patel
University of Texas at Arlington
fxp3176@mavs.uta.edu

Josue Caraballo
University of Texas at Arlington
josue.caraballo@mavs.uta.edu

Patrick Hennecke
University of Texas at Arlington
pwh8492@mavs.uta.edu

Chengkai Li
University of Texas at Arlington
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 tools only support data collection using simple keyword-based or account-based search. On the contrary, Wildfire employs a heuristic graph exploration method to selectively expand the collected tweet-account graph in order to further retrieve more task-relevant tweets and accounts. This approach allows for the collection of data to support complex social sensing tasks that cannot be met with a 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.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; • **Information systems** → *Data analytics*; • **Applied computing**;

KEYWORDS

social sensing; social media analytics; data collection; natural language processing

ACM Reference Format:
Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang, Foram Patel, Josue Caraballo, Patrick Hennecke, and Chengkai Li. 2024. Wildfire: A Twitter Social Sensing Platform for Layperson. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3616855.3635704

WSDM '24, March 4–8, 2024, Merida, Mexico

ACM ISBN 979-8-4007-0371-3/24/03.
https://doi.org/10.1145/3616855.3635704

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* [8], 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 can capture real-time changes in public opinion, setting it apart from traditional polls. Moreover, social sensing offers unique insights from data that only exist in social media, such as "likes." Furthermore, technological advancements in deep learning and natural language processing have greatly amplified the benefits of social sensing via sophisticated, large-scale analysis of social media data.

Many of those who need to conduct social sensing in the real world are non-technical users. However, social sensing requires complex skills in coding, data processing, and analytics. It is non-trivial to retrieve the data from social media platforms (e.g., through various APIs, using search and query conditions), to parse the data which could be in formats such as JSON and may contain both texts and multimedia, to store them under 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,[1] one often needs to devise sophisticated algorithmic approaches to prioritizing which part of the data to retrieve before others. Once the data are in place, various types of content analysis (e.g., sentiment analysis), 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.

There exist many different social sensing tools. Open-source tweet collection tools such as [3, 5, 7] utilize Twitter APIs and primarily offer basic data statistics such as hashtags and mention frequencies. Commercial social sensing tools (e.g., brandwatch.com, digimind.com, hootsuite.com, and synthesio.com) excel in comprehensive data analytics. However, they mainly focus on business analysis,

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

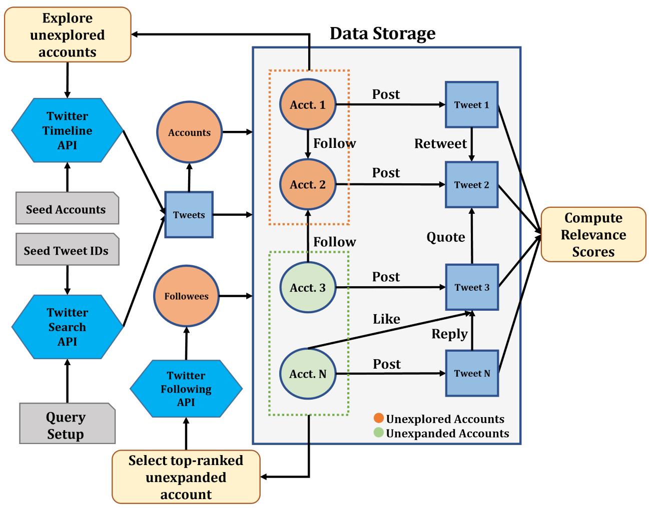

Figure 1: Data Collection Architecture

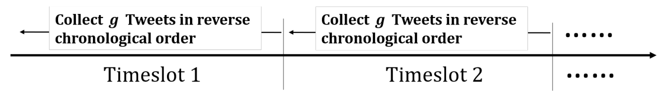

Figure 2: Timeslot and Granularity

and there are limitations on customization. Users may find it challenging to tailor the tool to their specific needs. For example, these tools do not support collecting data through graph exploration methods, such as community detection algorithms. Additionally, they do not provide users the freedom to leverage a variety of cutting-edge machine learning models according to users' needs.

This paper introduces Wildfire (demo and video tutorial available at https://idir.uta.edu/wildfire), a novel social sensing platform dedicated to Twitter data. It enables anyone with access to Twitter APIs to conduct social sensing tasks on large volumes of data without writing a single line of code. Our contributions include:

- Wildfire offers a flexible data collection mechanism, harnessed through Twitter's APIs, and a heuristic graph exploration approach. This method expands the initial seed collection of tweets and accounts by employing a ranking function to iteratively identify target accounts and tweets via "following" relationships. The ranking function is guided by a set of weighted classifiers, including those developed in-house for tasks such as detecting the check-worthiness of factual statements [4], as well as classification models using ChatGPT and from HuggingFace.
- Wildfire uses a single graph to store all Twitter accounts and tweets collected for multiple tasks. Meanwhile, collected data are annotated to discern the portion of the graph associated with each specific task. This approach enables Wildfire to avoid redundant collection and storage of data.
- Wildfire provides a range of analytic tools (e.g., text classification, topic generation, and entity recognition) which are crucial for accomplishing common tasks in social sensing such as finding trending topics and events.

2 SYSTEM DESIGN

Figure 1 shows the data collection architecture of Wildfire. The collected tweets and 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, users must sign up and provide a Twitter bearer token.[2] Once logged in, users can create multiple social sensing tasks. The system uses the user-provided Twitter token to collect data for these tasks. Wildfire supports concurrent execution of multiple tasks by the same user. Instead of populating a separate graph for each task, Wildfire stores data across all tasks 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 unique design allows Wildfire to share data across tasks, reducing users' token usage and shortening task completion time.

Given the aforementioned system design, the discussion in this section focuses on how Wildfire collects data. The system populates the graph with two methods using Twitter APIs as follows.

1) The *seed collection* method uses the Search API to obtain tweets and their corresponding Twitter accounts, based on user-specified keywords or Twitter queries.[3] Alternatively, the user can choose to provide IDs of seed tweets and seed accounts to be collected.

Since Twitter limits the number of historical tweets that can be retrieved each month and returns tweets in reverse chronological order, Wildfire requires the user to specify a few parameters when creating a search 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). Wildfire splits the search period $[s, e]$ into $\lceil \frac{e-s}{t} \rceil$ 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 denote accounts as *unexplored* if the system has not collected their timeline tweets, and as *unexpanded* if it has collected their timeline tweets but not their following lists.

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 relevant to the given task. The use of the ranking function follows the rationale that a relevant

[2] https://developer.twitter.com/en/docs/authentication/oauth-2-0/bearer-tokens
[3] https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

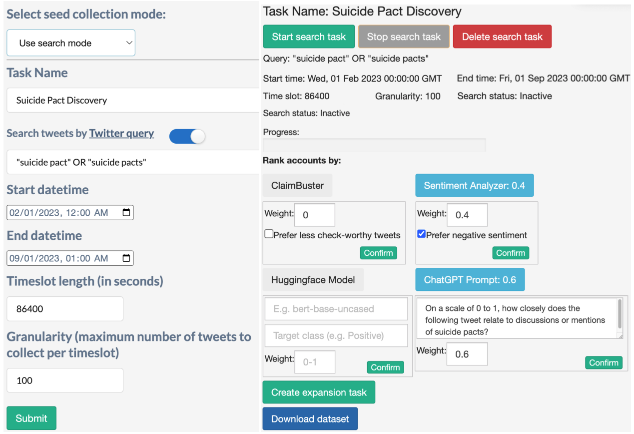

Figure 3: (a) Task Creation (b) Task Monitoring/Expansion

Twitter account's followees may be relevant too since they may share common interests and beliefs and they may participate in joint or similar conversations. This expansion collection can be run concurrently with the seed collection, thereby accelerating the data collection process.

Wildfire enables users to use multiple weighted classifiers to construct their ranking functions. The classifiers aim to identify tweets exhibiting specific attributes, such as particular sentiments and political affiliations. Each individual classifier, denoted C, calculates a relevance score, denoted $C(t)$, for a given tweet t. The relevance score for account a under a classifier, based on a collection of N_a timeline tweets for that account, can be determined by calculating the average relevance score across those N_a tweets. Users have the flexibility to specify a customized ranking function by assigning a weight (w) to each of these account relevance scores. Users also have the option to score tweets and accounts in ascending order instead of descending order if lower values align with their interests, e.g., a user wishing to collect tweets with negative sentiment would prefer lower score values. Equation 1 shows how the ranking function selects the top-ranked account from the pool of unexpanded accounts (X) using k classifiers, as follows.

$$Top(X) = \underset{a \in X}{argmax} \left(\sum_{i}^{k} w_i * \frac{\sum_{t \in a} C_i(t)}{N_a} \right) \tag{1}$$

Currently, Wildfire provides 2 built-in classifiers: *ClaimBuster* [4], a deep learning model to detect tweets containing check-worthy factual claims, and *RoBERTa sentiment analyzer* [1], a multilingual model to discover the sentiment of tweets. Besides them, users can also utilize any classifier from HuggingFace by specifying the model name and the desired class (e.g., preferring negative sentiment). Wildfire employs these classifiers through HuggingFace Inference API calls, providing users with a diverse array of models to select from. Furthermore, Wildfire provides users with the option to employ ChatGPT as a classifier by simply providing a prompt that leads to relevance scores from ChatGPT.

3 DEMONSTRATION

Wildfire's user interface comprises three main components: 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 enables users to set up a task with various settings. The task monitoring and expansion page displays the status and progress of tasks, allowing users to control each task and configure its expansion. The data analytics page of each task allows users to explore collected tweets using multiple filters and examine corresponding analytics results.

Task creation (Figure 3(a)) Users have the option to create a task within one of three settings: search mode, tweet ID mode, or account mode. For the search mode, users can further toggle between simple (i.e., keywords) and advanced search (i.e., query). The simple search allows users to enter multiple keywords separated by commas. The advanced search allows users to enter a Twitter query. [3] If users already possess specific tweets or Twitter accounts of interest and intend to gather additional tweets using the ranking function, they can choose either tweet ID mode or account mode, where they must specify the tweet IDs or Twitter account handles. Users can also specify *Start datetime*, *End datetime*, *Timeslot length*, and *Granularity*, as discussed in Section 2.

Task monitoring and expansion (Figure 3(b)) This page provides the overview of each task, displaying its configuration, status (such as active, completed, stopped, and error), and collection progress. For creating the task expansion, users can choose the weight and preference associated with each ranking factor. Users are given options to start, stop, resume, or delete both seed and expansion collection tasks. The "Download" button at the bottom of each task leads to a page where users can choose what to include in the downloaded dataset, e.g., whether to include the profile of each Twitter account.

Data analytics (Figure 4) 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, Twitter account handles, date range, and classification score range. Once the desired filters are applied, the corresponding result tweets are displayed and paginated. The analytic tools include topic generation for generating thematic tags of tweets, entity recognition for extracting entities from tweets, and the aforementioned two classifiers. Users have two ways to obtain analytic results. The first is to click 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 two classification score distributions presented as a pie chart, a bar chart displaying the top 10 most frequent tags representing the topics, and another bar chart showing the top 10 most frequent entities recognized from the tweets. The second way is to click any tweet on the page, which opens a pop-up window (orange frame in Figure 4) showing the results of all four tools applied to the tweet. The output of the two classifiers is presented as scores between 0 and 1. The topic generation tool generates a list of topic words derived from the tweet text, although these words do not necessarily appear in the tweet. The results of entity recognition are provided as pairs of entities and their corresponding entity types.

Method	Total collected	Sample size	Accuracy
Seed collection	4,347	100	48%
Expansion collection	6,280	100	67%

Table 1: Comparison of Task Relevance in Seed and Expansion Collections

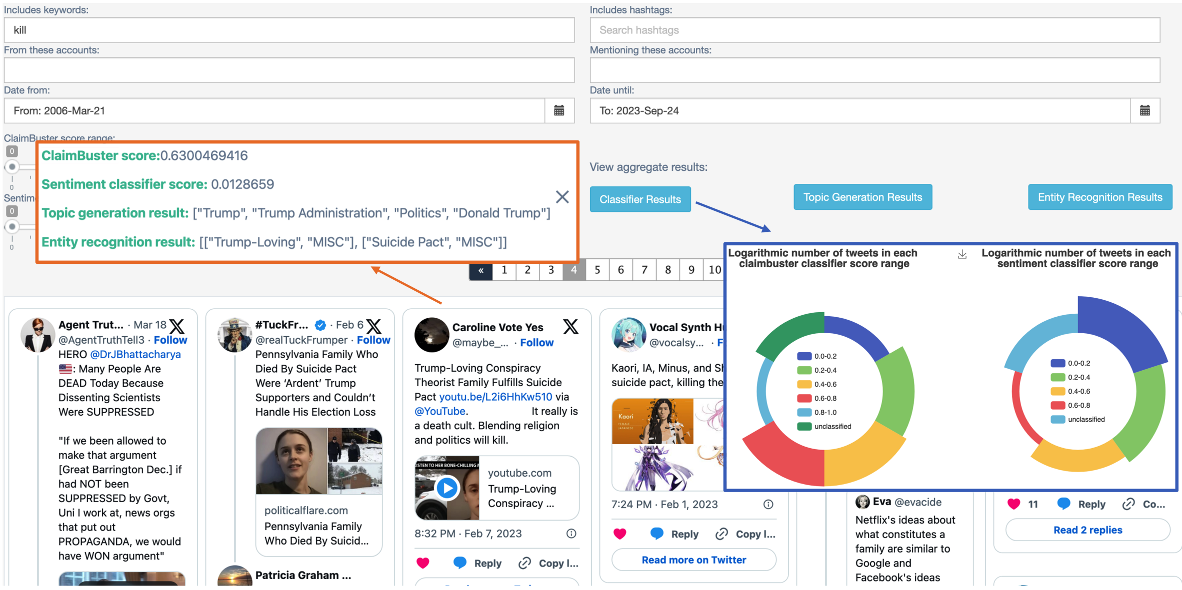

Figure 4: Data Analytics Page

4 EXPERIMENTS

Many social sensing studies involve time-consuming data collection and analysis processes. It is high time to have an automated tool to aid such studies. Therefore, by presenting two use cases that replicate existing studies, we demonstrate the effectiveness of Wildfire in facilitating social sensing studies.

Use Case 1: Identifying political polarization on social media. To study political affiliations (e.g., left-wing or right-wing) on social media, researchers often use keywords or hashtags to collect tweets and accounts in favor of or against certain political affiliations. For instance, [2] examined political polarization during the 2016 U.S. presidential election. The study used a dataset of millions of tweets collected by keywords and hashtags that reflect support for two presidential candidates. Wildfire can help improve the data collection and analysis. We created a seed collection using keywords from the paper and retrieved 4,347 tweets. By selecting the HuggingFace model politicalTweetBERT in the ranking function, Wildfire found more tweets expressing support for Democrats. The expansion collection brought an additional 6,280 tweets. By comparing two samples of 100 tweets drawn from the seed collection and the expansion collection respectively, we found the high-scored tweets from the expansion were more relevant (67%) than those from the seed collection (48%), as illustrated by Table 1).

Use Case 2: How people make suicide pacts on Twitter. To investigate whether individuals utilize Twitter to search for like-minded persons and form suicide pacts, [6] collected tweets that contain "suicide pacts" in Korean from Oct. 16th to Nov. 30th, 2017. We used Wildfire to create a seed collection with the same keywords in English and incorporated an expansion that applies Sentiment Analyzer and ChatGPT with the prompt "On a scale of 0 to 1, how closely does the following tweet relate to discussions or mentions of suicide pacts?" in the ranking function. The seed collection retrieved 10,953 tweets, and within one hour of its expansion, an additional 24,112 tweets and identified 8 accounts were retrieved as well. This data spans from February 1st, 2023, to March 24th, 2023. Among the tweets from the expansion collection, 411 tweets are scored above 0.9 by ChatGPT. By manually checking 100-tweet samples from the 411 high-score tweets, we found that 79 tweets from 3 accounts are relevant to suicide. This indicates that the expansion equipped with ChatGPT (Figure 3(b)) effectively identified the target tweets.

5 FUTURE WORK

Recent revisions to Twitter's API policies have brought a significant escalation in the expense of accessing Twitter data. For example, the free API access for academia has ceased. Instead, anyone seeking Twitter data is now required to pay, with prices considerably higher than previous rates (e.g., $5,000 for 1 million tweets per month). We are diversifying our data sources by integrating alternatives such as CrowdTangle, thus reducing our reliance on Twitter.

ACKNOWLEDGEMENT

This work was partly supported by the National Science Foundation under Grants 1937143.

REFERENCES

[1] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 258–266. https://aclanthology.org/2022.lrec-1.27
[2] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. 2020. Learning political polarization on social media using neural networks. *IEEE Access* 8 (2020), 47177–47187.
[3] 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.
[4] 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.
[5] Michael W Kearney. 2019. rtweet: Collecting and analyzing Twitter data. *Journal of open source software* 4, 42 (2019), 1829.
[6] Sang Yup Lee and Yeji Kwon. 2018. Twitter as a place where people meet to make suicide pacts. *Public Health* 159 (2018), 21–26.
[7] George Washington University Libraries. 2016. Social Feed Manager. (2016).
[8] Dong Wang, Boleslaw K. Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. 2019. The Age of Social Sensing. *Computer* 52, 1 (2019), 36–45.