Using Social Media To Predict the Future: A Systematic Literature Review

Lawrence Phillips[1*, ☯], Chase Dowling[1, 2, ☯], Kyle Shaffer[1], Nathan Hodas[1], Svitlana Volkova[1]

1 Data Sciences and Analytics Group, Pacific Northwest National Laboratory, Richland, Washington, United States of America
2 Electrical Engineering, University of Washington, Seattle, Washington, United States of America

☯These authors contributed equally to this work.
* Lawrence.Phillips@pnnl.gov

Abstract

Social media (SM) data provides a vast record of humanity's everyday thoughts, feelings, and actions at a resolution previously unimaginable. Because user behavior on SM is a reflection of events in the real world, researchers have realized they can use SM in order to forecast, making predictions about the future. The advantage of SM data is its relative ease of acquisition, large quantity, and ability to capture socially relevant information, which may be difficult to gather from other data sources. Promising results exist across a wide variety of domains, but one will find little consensus regarding best practices in either methodology or evaluation. In this systematic review, we examine relevant literature over the past decade, tabulate mixed results across a number of scientific disciplines, and identify common pitfalls and best practices. We find that SM forecasting is limited by data biases, noisy data, lack of generalizable results, a lack of domain-specific theory, and underlying complexity in many prediction tasks. But despite these shortcomings, recurring findings and promising results continue to galvanize researchers and demand continued investigation. Based on the existing literature, we identify research practices which lead to success, citing specific examples in each case and making recommendations for best practices. These recommendations will help researchers take advantage of the exciting possibilities offered by SM platforms.

Introduction

"Forecasting is a difficult business, particularly when it is about the future."

– Yogi Berra

Now more than ever before, companies, governments, and researchers can gather and access data about people on a massive scale. Putting a finger on the pulse of public opinion is made increasingly possible thanks to the rise of social media (SM; for a more comprehensive review of SM platforms see [15, 107]). SM are Internet-enabled platforms that provide users with a persistent online identity and means of sharing information with friends, families, coworkers, and other users. There are many different SM platforms, each of which targets a different aspect of what users want or need: e.g., LinkedIn targets professional networking activities, Facebook provides a means of connecting friends and family, and Twitter provides a platform from which to quickly broadcast thoughts and ideas. These platforms are incredibly popular: as of June 2016, Facebook sees an average of 1.13 billion daily users, including nearly half the populations of the United States [24] and Canada [28] logging in every day [70].

Being so widely used, SM platforms generate massive quantities of data. According to [106], in 2013 users were posting an average of over 500 *million* tweets every day. While traditional data sources and records of daily human activity, such as newspapers and broadcast media, are often constrained by national, cultural, and linguistic boundaries, SM platforms are generally consistent provided a user has access to the Internet. Moreover, traditional media requires time to compile relevant information for publication, while SM data is generated in real time as events take place.

All of this information can be collected and mined by virtually anyone who wishes to use it. As far back as 2009, the United States Geological Survey (USGS) began investigating the possibility of using SM data to detect earthquakes in real time [69]. Information about an earthquake spreads faster on SM than the earthquake itself can spread through the crust of the Earth [104]! Similarly exciting work in forecasting with SM also exists; EMBERS is a currently deployed system for monitoring civil unrest and forecasting events such as riots and protests [162]. Using a combination of SM and publicly-available, non-SM data, they are able to predict not just when and where a protest will take place, but also why a protest may occur. These findings have enticed researchers into exploring the possibilities opened by SM data, but there remain many unanswered questions. If SM is useful for detecting real-time events, can it be used to make predictions about the future? What limitations does forecasting with SM data face? What methods lead researchers to positive results with SM data?

For all of its exciting advantages—SM platforms are global, multilingual, and cross-cultural—a deep pessimism surrounds SM data analysis [167, 204]. SM is noisy and the data derived from SM are of mixed quality: for every relevant post there may be millions that should be ignored. Learning with SM data sometimes requires robust statistical models capable of handling massive quantities of SM data, but here too there are additional open questions about the effectiveness of such data-driven models. Consider the case of Facebook, who in 2014 launched a *trending topics* feature later revealed to be hand-curated by Facebook employees [160]. Facebook used an algorithm to scour the site, utilizing their own SM platform's data to detect trending topics that were then looked over by humans for quality assurance. Facebook later removed human curators from the process—following the idealized trend of SM data analysis—and relied entirely on their data-driven algorithms. Within days Facebook's system had posted libelous articles and explicit material [187]. If a SM platform as large as Facebook is unable to use its own data to detect aberrant trending topics, what are the prospects for other organizations?

Yet, in spite of anecdotes like this, researchers continue to investigate how best to make use of SM data. Preliminary results do largely show positive findings as we discuss in much greater detail below. If SM users are reacting to and talking about events in real time, one might imagine that users are also talking about and reacting to events that they anticipate will happen in the future. This raises the interesting possibility that SM data might be useful for *forecasting* events: making predictions about events that have yet to occur. Not only have researchers begun to investigate this line of questioning, earlier review articles on SM forecasting showcase early positive examples of predictive success [100, 146, 176, 216]. Across the board preliminary studies show that SM *could* be used to predict the future. At the same time, early findings have been controversial and warrant some amount of skepticism and caution [100, 176, 216]. The field is in its infancy, methodologies are scattered, common best practices are nonexistent, and true replication of studies is near-impossible due to data sharing concerns [204].

Previous reviews laid out a number of possible issues with SM forecasting and identified areas where forecasting had or had not been successful, but had little to say about what best practices researchers might follow in order to better make use of SM data. Identifying all of the pitfalls associated with SM data is far beyond the scope of this literature review, therefore we choose to focus on the following general questions:

Q1: Can SM be used to make accurate predictions about current and future events?

Q2: Across domains, what distinguishes SM prediction successes from prediction failures?

While previous reviews were cautiously optimistic in addressing Q1, by covering a much larger

body of literature, we aim to find a more comprehensive answer. We further address Q2 in order to give researchers an idea of how they might best approach their own SM forecasting tasks. The contents of the rest of this review are organized as follows: the background section provides a general overview of SM and the interest it has generated, clarifies the meaning of *prediction* and *forecasting*, and describes some general challenges faced by researchers. We describe which general topics are covered in the literature review and methods and requirements for study inclusion. Next, we present our findings split by prediction topic, focusing on elections, economics, public health, threat detection, and user characteristics, addressing research questions (Q1) and (Q2) above. Further, each results section includes a table of reviewed articles that lists the primary author, topic, data source, collection method, size, primary data features, algorithmic task, success rate, and validation method of the section's constituent reviewed articles.

Our principle research questions (Q1) and (Q2) relate to how well SM data can be used to predict future (or otherwise unknown) real-world states, i.e., *forecasting*. We also note that many papers focus on identifying the current state of the world, i.e., *nowcasting*. Both types of papers are included in our analysis for two principle reasons. First, the state of the world is often persistent over time, meaning that current predictions may overlap with future predictions, e.g., the case of predicting a user's ethnicity. Second, predicting the future is likely to be more difficult than predicting current states because of increased temporal distance [12]. The ability (or inability) of existing research to nowcast current or immediate future states is therefore an upper bound on how well forecasting states further in the future might perform.

Background

The use of SM data for modeling real-world events and behavior has seen increased interest since its early appearances in academic work around 2008. Fig 1 illustrates this growth, with nearly 20,000 articles having been published in 2015; meanwhile, 2016 is set to well exceed that number. This rise in popularity is commensurate with the newly coalescing field of computational social science [112]. Many sociological hypotheses were previously untestable due to difficulties in obtaining data. With the advent of SM, this is no longer the case, as myriad facets of human interaction are recorded by millions of people across the web. At the same time, this data is not always a complete cross section of what a researcher might hope to see. SM usage varies by age, culture, socioeconomic status, gender, and race [155]. Still, positive findings and interest in the fundamental dynamics of SM platforms is a likely culprit for this exponential growth in popularity, particularly for social scientists [15, 79, 86, 107, 146, 204, 218].

Forecasting and Predictive Modeling

Standard examples of physical laws and theories (e.g., Newton's Laws or the Ideal Gas Law) have provided the sciences with a means of *forecasting* or *predicting* natural phenomena. Specifically, given a sequence of observations related to the state of some system, *prediction* entails the accurate and reproducible state estimation of that system for some amount of time into the future up to and including the present. For a simple physical system, we might use Newton's Laws to derive a model of the position and velocity for a mass on a spring (i.e., Hooke's Law). Models which build off a theoretical understanding of the underlying system are considered *theory-driven* models. In many cases, however, we lack a full or even partial theoretical understanding of the underlying system. For instance, it would be quite difficult to create an entirely theory-driven model to forecast when a user is going to make their next SM post about an unforeseeable topic. *Data-driven* models learn predictive relationships from data directly, for instance by looking at previous posting patterns for a user. We distinguish between theory- and data-driven models, although in practice models often incorporate aspects of both methods. Data-driven models are often used to gain insight into the fundamental laws governing the underlying system: the authors

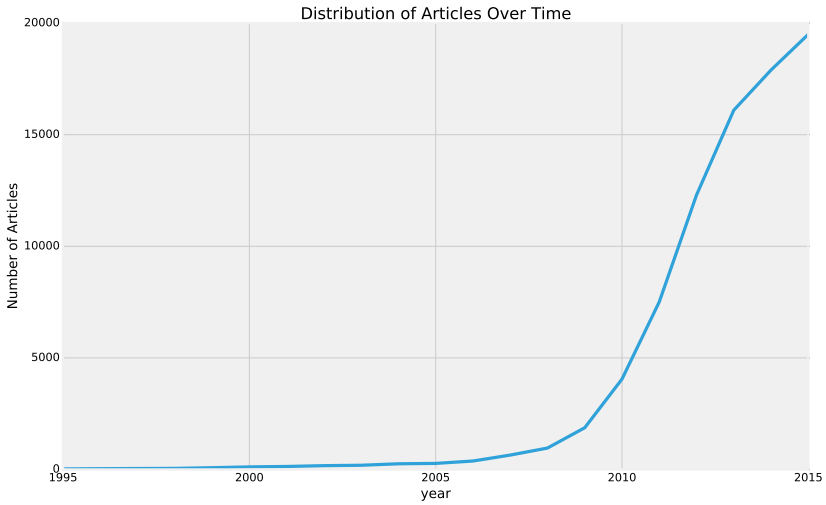


Fig 1. Number of articles published per year containing the phrase "social media" and the keywords "data" and "prediction" according to Google Scholar, excluding patents and case law.

of [189] demonstrate how to recover or learn Hooke's Law directly from sensor data without knowledge of Newton's Laws.

Underlying Complexity of SM-based Models

It is clear that forecasting should be possible to varying degrees when there are direct causal links, as in the case of the physical systems described above, whether these links are identified through theory-driven hypothesis testing, naive data analysis, or both. If current weather patterns impact future weather, that relationship should allow for forecasting. If current behaviors impact future chance of illness, that relationship should likewise allow for forecasting. Yet in almost all cases, SM posting does not directly impact the real world system we care about and the real world system does not directly impact SM posts or behavior (which is generally the relationship being modeled for SM forecasting). Instead, physical systems in the real world and SM users interact with one another and then users interact with SM. We demonstrate these relations in Fig 2. Each arrow represents a (not necessarily causal) relationship between two systems which can be modeled, where direction matters. Forecasting can be accomplished by using theoretical knowledge to understand the underlying mechanisms which produce a link between SM behaviors and real-world outcomes, or this relationship can be modeled directly from the data.

Prediction becomes somewhat more difficult as the gap between any two of these factors increases and their relationships becomes less *direct* between one another. While many users may be influenced by or be an influencer of a stock market, for example, predicting the behavior of a stock market is already known to be an all but insurmountable task both empirically [14, 123] and theoretically [127]. While focused SM data analysis may yield insight into stock market behavior, SM users (corporate or otherwise) are unrepresentative of the players within a stock market and trades are often purposefully obfuscated [30]. This is to say that SM does not significantly overlap or impact a majority of the variables governing the physical system, namely a stock market in this case. The difficulty of establishing SM prediction for real-world events is a reflection of these underlying processes which vary between tasks and are often only poorly understood or not taken into account.

A significant manifestation of an event on SM, however, does not appear to be a sufficient condition for successful prediction. Take for instance the 2014 World Cup; the tournament saw global SM presence representing participating teams from around the world [48, 62, 81, 161, 217].

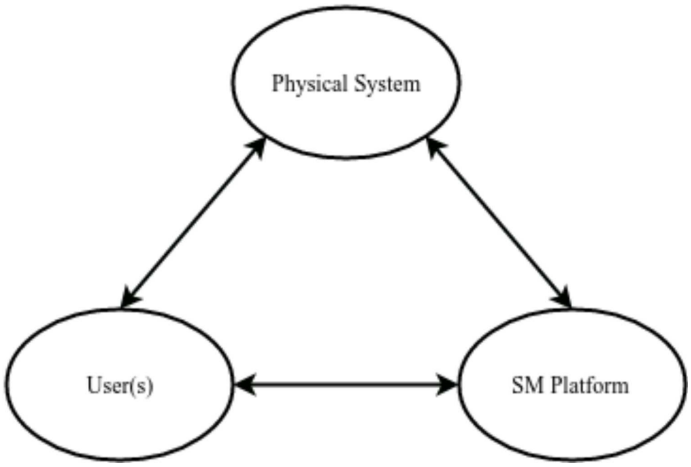


Fig 2. Interactivity of factors leading to SM forecasting ability. For example, users may observe the weather around them and post those observations on Twitter. A less common example would be weather directly effecting the social media platform, via weather-related outages.

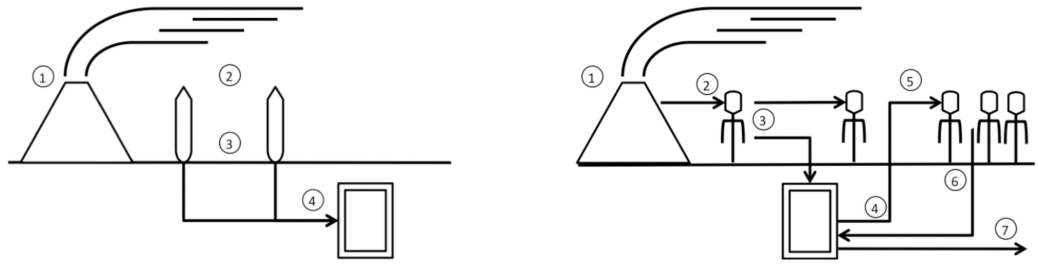
An attempt to predict match outcomes utilizing Twitter data, [161] failed to perform better than random chance for early tournament matches, and under-performed popular sports analysis agencies' predictions beyond quarter-final matches. Much of any given team's SM traffic reflected the development of a game and general attempts to rally fan pride [217], but the SM platform's activity itself had little demonstrable bearing on the outcome of the game. Indeed, apart from a handful of sports journalists broadcasting informed *a priori* analysis, the majority of fans are not directly involved in the game and are merely spectators possibly explaining poor predictive performance [161]. On the other hand, as spectators, SM users do post information which can be used to identify what is happening as a match progresses, i.e. nowcasting [48, 133].

Ultimately, additional variables in complex prediction tasks increase the gap between each of the factors in Fig 2. Simpler queries with direct relevance to how users interact with SM and the physical system might be expected to enjoy better predictive success. Consider the case of predicting when soccer matches like the above will occur. The authors of [93] achieve an accuracy of $\pm$ 8 hours up to 6 days in advance of a game. This could be attributed to the fact that attending a game directly impacts all users involved—players, journalists, and fans alike—where fans will broadcast their planned attendance and support on SM in addition to teams, players, and journalists publicizing the event [48]. Additionally, in some cases SM users have *direct* knowledge related to the forecasting task, e.g., they do know when a game will take place well ahead of time. Such instances should be much easier to forecast than cases where any knowledge on the part of SM users is *indirect*, as in forecasting match winners, where it could be argued that SM users are privy to some relevant information, but have no direct knowledge of the outcome.

The SM User-Sensor

Users themselves further complicate matters. Consider the process for detecting an event with a traditional sensor network in Fig 3a taken from [47]: 1) some physical event occurs, 2) sensors acquire a measurement, 3) the sensors record the measurement, and 4) the system stores the measurement. Although sensor readings may be correlated, the sensors do not typically interact with one another directly. SM users can be thought of as sensors, but the purpose of the SM platform is specifically to *allow* interaction between different users. Consider the parallel process of event responses in a SM sensor network in Fig 3b: 1) some physical event occurs, 2) user receives stimulus, 3) user communicates response, 4) system routes message, 5) other users receive message, 6) users communicate response, and 7) system routes message.

Consider an idealized case of a traditional sensor network where one sensor is reporting false



(a) A traditional sensor network **(b)** A SM sensor network

Fig 3. Comparison of sensor data routing: traditional vs SM sensor network, adapted from [47].

information. In such a case, the incorrect sensor's data can be compared against the data received from other sensors and because the sensors do not interact with one another, a single incorrect sensor will not cause a cascade of false information. On SM, however, such information cascades can and do occur. Consider the case of the 2013 Boston Marathon bombing. Immediately following the event, users on various SM platforms, in particular Reddit, began an attempt to identify the bombers. As SM users shared information with one another they mistakenly settled on Sunil Tripathi as the primary suspect. Tripathi had been missing for a month by the time the bombing took place and had in fact taken his own life [198]. Because SM users react not just to outside events but also to posts from other SM users, it is possible that a user's perception of outside events is influenced by other users, essentially introducing the possibility of sensors biasing other sensors. Besides false accusations, this leaves SM sensors susceptible to other well-studied phenomena such as group polarization [44, 213].

Open Challenges for SM Forecasting

This is all to say that while SM data holds tremendous potential value, its useful application is not necessarily a trivial matter. Forecasting techniques in the natural sciences, both theory- and data-driven, are relevant, but SM challenges researchers to find new ways to apply them. Aggregation techniques from traditional sensor networks are relevant, but SM challenges researchers to find new ways to augment them. Because of these difficulties, Q1 and Q2 are intricately linked. To be able to generate valid, reliable predictions (Q1) researchers must first identify the methods through which myriad challenges in SM research may be addressed (Q2). These difficulties include noisy data, possible biases, a rapidly-shifting SM landscape which impedes generalizability, and the need for domain-specific theory to wrap everything together. In order to address whether these challenges can be overcome, it is necessary to examine the literature in a systematic fashion.

Methods

In this section we detail the methods used in our systematic literature review. We define the scope of the review, describe how studies were collected and reviewed for inclusion, and discuss potential sources of study bias.

Task Overview

While many researchers have acknowledged the potential usefulness of collecting and analyzing SM data as a way to study social phenomena, much past work has concentrated on predicting various *online* aspects of social networks, in particular virality [205] and information cascades (message propagation) [185]. We restrict ourselves to reviewing work that focuses on using *online* data to predict *offline*—viz. physical world—events. We refer to these as 'real-world phenomena.'

Previous reviews have covered similar ground but describe results without clearly identifying what aspects of each domain or methodology led to success or failure of SM prediction [15, 100, 176, 216]. For instance, [100] reviews the literature in 2013, and makes some very general statements regarding what techniques lead SM papers to demonstrate successful results. Unfortunately, the authors collapse these generalizations across all research domains, making it difficult to discern what techniques might be best applied in particular disciplines. A taxonomy of predictive models used in SM research is provided by [176] and explores more specific issues by content domain, but looks at specific case-in-point examples: influenza, product sales, stock market, and electoral predictions. In a like manner [216] covers a small set of content areas and also is unable to draw strong conclusions. All three reviews come to the basic conclusion that SM should be able to make accurate predictions about current and future real-world events (Q1), but are either somewhat pessimistic or unclear about how this might be feasibly accomplished. Because of the limited scope of previous reviews, no one in the literature has adequately addressed our second primary research question (Q2): what distinguishes success from failure in studies of SM prediction across all domains?

As can be seen from Fig 1, a great deal of research on SM prediction has been published since the last round of literature reviews in 2013. We take advantage of this greater body of literature in order to expand our review, drawing more specific conclusions about what leads to successful predictions. In order to understand how SM prediction functions both within and across domains, we divide this review on the basis of previously well-trodden disciplines. These disciplines represent the most active research areas where SM data is being used to predict real-world phenomena.

We first provide a general outlook for each discipline as well as the specific types of prediction tasks for which researchers in these areas use SM data. We present a table of articles including the primary author, topic, data source, collection method, size, primary data features, algorithmic task, success rate, and validation method of the section's constituent reviewed articles. We then discuss general findings from each paper, noting in particular which specific factors appear to either reduce or increase the success found by researchers in that domain (Q2). Finally, we compare the existing literature in each field across disciplines to identify which methodologies are most promising by demarcating specific examples of successful and unsuccessful research practices.

Data Collection

To perform a systematic literature review in a research area making rapid advancements, it is important to review all topic-relevant papers regardless of their place of publication. We follow the guidelines set by PRISMA with our search outline described in Fig 4 [138]. We first conducted a database search in October 2016 of IEEE and ACM for articles published since 2010 which included the terms"social media" and "prediction" as well as for "forecasting" in the case of IEEE. This returned a total of 905 search records. We augmented this number by backward tracking references from previous literature reviews on SM forecasting as well as by searching articles from the following conference proceedings and their associated workshops from 2010 through August 2016: ACL, EMNLP, EACL, NAACL, WWW, KDD, NIPS, WSDM, ICWSM, CHI, ASONAM, AAAI, IJCAI, and SocInfo. This resulted in an additional 86 records.

After removing duplicates this left 958 abstracts to screen. We included only those articles which attempted to use SM data to make real-world predictions. This included making predictions about the state of the world currently, which we refer to as "nowcasting", or making predictions

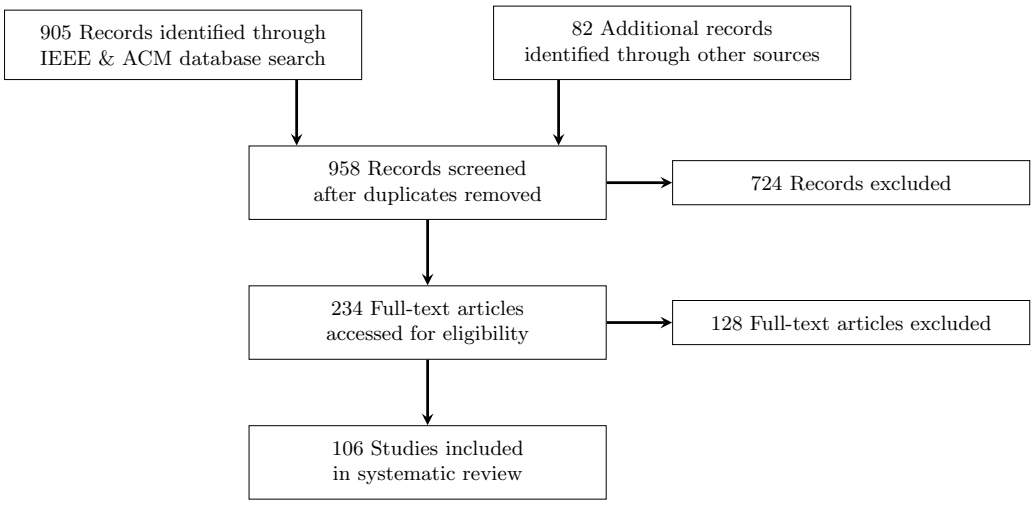


Fig 4. PRISMA Flow Diagram: 958 abstracts were gathered through a database search of IEEE and ACM along with a search of relevant conference proceedings. Of those, 234 full-text articles were screened for eligibility resulting in a final set of 106 studies included in the final systematic review.

about future states of the world, which we refer to as "forecasting". Articles were excluded if they did not either make predictions or attempt to discover relationships with real-world events or characteristics, e.g., speculative or theoretical articles. We purposefully excluded all articles which use SM data only to predict future SM data, e.g., research on "virality" which predicts the spread of SM posts on SM platforms. After the abstract screening process, this left 234 full-text articles which needed to be assessed for inclusion. Articles were excluded for any of the above reasons as well as for failing to report concrete quantitative results, lacking real-world ground truth data, being primarily a review article, or possessing serious and obvious methodological concerns. This left a remaining 106 articles for inclusion in the systematic literature review.

For each of the examined full-text articles, we collected information regarding the authors, topic of the study, SM platform(s) analyzed, data size (e.g. number of users, posts, images), SM features used in their analyses, the type of prediction task (e.g. classification, regression), their principle success metric and results. Because of privacy concerns, data collection methods are not always published in full and therefore where data size was not made publicly available this is noted. SM features were classified into a number of discrete categories including user metadata, n-gram counts, semantic (NLP) features, social network features, spatial or geolocation features, post volume, user behavioral features, and other non-SM features. Where multiple evaluation metrics were reported, we focus on those results primarily highlighted by the author(s) or which best represent the best level of performance achieved.

Study Bias

For researchers hoping to make use of SM data for their own prediction tasks, we must qualify our observations by noting that we are unable to provide a systematic analysis of work which has not been published. Bias to publish studies with positive results necessarily taints our view of what SM can accomplish [167]. There may well be domains where SM forecasting has been attempted, failed, and the results were not published. Because of our ignorance in these matters, we can make only reasoned assumptions about the possibility of success in domains not represented in the current review.

Further, the selection of current studies is biased in terms of which SM platforms have been studied. By far the most studied platform has been Twitter, due in large part to the ease of

acquiring its data [146]. Much of the research on Twitter may not be applicable to alternative SM platforms. Although images play a crucial role in SM they are particularly understudied and therefore we can say little about their possible predictive value [5, 98, 158, 208, 215].

Likewise, current research has largely focused on SM data in English and on events in the United States. It is unclear how well techniques suited toward the demographics of the U.S. can be applied to other countries, although where this has been explicitly conducted there have been largely positive results [59, 172, 188].

Results

Research on SM forecasting spans a very wide range of topics. In order to make better sense of the existing body of literature, we split our discussion based on five general domains which have been most explored: Elections and politics, stocks and marketing, public health, threat detection, and user characteristics. For each of these domains, we discuss the existing literature in terms of its general topics and methodologies, noting particular successes and failures. We also present a detailed table which includes a number of characteristics for each study including the article's topic, data source and size, features used, the type of task (classification or regression), and their reported results.

Elections

Research in election prediction has provided significant insight into the capabilities and limitations of predictive models trained using SM data. Social media platforms have allowed users to share their opinions and sentiments on a variety of topics, particularly in political discourse, and this has spurred a great deal of interest in predicting the outcomes of elections and other policy issues [44]. Political forecasting is one of the first content areas to be explored with SM data, with a number of studies published by 2010 [82, 144, 191, 206], with a comprehensive meta-analysis conducted in 2013 [78]. Given that these platforms provide a large archive of how people have talked about political and social issues, researchers have investigated the utility of this potentially useful data source in predicting and forecasting various aspects of elections and political life. In particular, research has largely focused on two specific tasks, forecasting election outcomes and now- or forecasting public opinion.

Election Outcomes

Many polling companies spend large sums to predict the outcomes of major elections. A great deal of early work on SM forecasting focused on the predictive power of microblogs, such as Twitter, to supplement or even replace expensive polling methods. Researchers have investigated a variety of techniques ranging from extremely simplistic [181, 191, 206] to somewhat more complex [29, 129]. Election forecasting is a very difficult task in part because major elections do not occur very frequently. Because SM is a relatively recent invention, training can only occur for a very limited number of past election cycles which may bias forecasting methods. Additionally, if a major election occurs only once per few years (e.g., 4 years in the case of U.S. presidential elections), then even legitimate predictors of a past election may well have changed in the intervening years.

The simplest method for forecasting election outcomes is based on assuming the volume of tweets mentioning a party or candidate reflects the share of the vote that will be won [102, 181, 191, 206]. These models collect tweets over a period of time before an election and filter for those tweets which mention a single candidate or party running for office. It is assumed that if a candidate is mentioned in 55% of these tweets, that they will receive 55% of the vote and will therefore be the winner. Two of the papers using this method purport to find extremely promising results [191, 206]. In particular, [191] report that their method for predicting German election outcomes is almost comparable to traditional polls. Unfortunately, later work has cast

much doubt on such simple methods. [99] replicate the work of [191] showing that the model relies crucially on excluding the German Pirate Party, a new party which represented 34.8% of mentions (almost twice that of the next most mentioned party, the CDU) but which garnered only 2.1% of the vote. Further, they show that even when excluding the Pirate Party slight changes to the dates of data collection can lead to major changes in forecasting error. A further difficulty for volume-based approaches is mentioned by [191], the 3.9% of users who tweet most heavily account for 44.3% of all political tweets in their data. Despite the possibility of heavy bias, they make no attempt to correct for this. Poorer performance for the same method is reported by [181] and [102] who forecast Singaporean and Indian elections, respectively. While Tumasjan reports a MAE of only 1.65%, the same method achieves 5.23% MAE [181] and 4.5% [102] casting further doubt on the utility of raw volume analyses.

One method for improving volume-based approaches is to take into account whether a candidate or party is mentioned in a positive or negative light. A number of studies explore whether this type of basic sentiment analysis might improve a simple volume-based approach, finding mixed results [38, 172]. For instance, [172] attempted to augment a basic volume-based approach by weighting total tweet counts based off the percentage of tweets which positively mentioned a political party. When combined with other normalization techniques (e.g., counting only one tweet per user and throwing away tweets mentioning multiple parties) sentiment improves results. Unfortunately, even after applying sentiment, the results of normalization are still worse than the basic predictions made by simple counting of mentions. The work of [38] models U.S. Republican presidential primaries from 2012 and counts a twitter user's vote as a function of the number of positive and negative tweets mentioning a candidate. Although they do not present results for a raw count prediction, their sentiment predictions are not particularly impressive. Broken down by user demographics, most groups struggle to reach above 50% (i.e., random) accuracy. Looking across various data collection time windows, only a single group (right-leaning Republicans) averages well above 50%, but even then stands at 67.5% averaged accuracy. The 2012 Republican primaries were also considered by [156] who examined the difference between traditional raw volume analyses and sentiment analyses which take into account the popularity of a post. They correlate predictions from blogs, Facebook, Twitter, and Youtube against traditional Gallup polls and present two major findings. First, blogs and Facebook provided strong polling forecasts when taking into account a post's popularity. In contrast, predictions from Twitter are much poorer overall and actually decrease when taking into account retweets while Youtube predictions are the worst quality regardless. Second, they correlate their forecasts against vote totals for each candidate replicating the above findings, with Facebook and blog posts being better real-world predictors. With such a limited number of data points to evaluate against, however, it's unclear whether these positive results might be replicated elsewhere. Poor findings in the field of election prediction overall suggest that volume of SM posts alone, with or without sentiment analysis, is likely a poor method for predicting election outcomes.

Prediction based solely on the number of tweets mentioning a candidate is a very rough method which fails to take into account a variety of other features which might be useful in predicting election outcomes. For instance, [27] demonstrates that while the number of Facebook friends a candidate has on election day correctly predicts winners in only 16.7% of the 2011 NZ elections, a baseline model featuring whether the candidate is an incumbent, whether the friends are of the same party as the incumbent, and similar control variables achieves accuracy of 88.1%. Adding Facebook predictions to these control variables improves accuracy to 90.5%. The work of [29] explores the possibility of modeling election outcomes based on tweet texts. They use a small set of hand-annotated tweets in order to estimate aggregate (rather than tweet) level sentiments. Using this method they are able to forecast the outcome of the 2012 French presidential elections roughly on par with traditional polling. They also forecast 2012 French legislative elections with a mean absolute error of 2.38 percentage points, as compared to an average of 1.23% for traditional polls. In predicting Taiwanese elections, [200] attempt to incorporate some notion of the popularity of SM posts. While this almost halves their prediction error, their results are still poor

with an MAE of 4.0%. Finally, [61] use the social graph structure of Twitter to forecast national and EU elections in Sweden. They restrict their analysis to the accounts of politicians with the idea that politicians should be more likely to win if they are influential in the SM graph, but report relatively poor correlations (EU $r = 0.79$; National $r = 0.65$).

From the existing literature it is clear that elections can be forecasted using SM data, although not with the same accuracy as traditional polling [29, 191]. Additionally, the types of features used play a large role in results. Simple volume is a very poor predictor [99] even when augmented with sentiment and taking into account the number of users rather than raw tweets [172]. The value of SM above and beyond simple baselines, however, may be relatively small [27] unless more advanced techniques can be utilized [29].

Public Opinion

An alternative goal for SM researchers has been to use online sentiment to nowcast public opinion, often with the goal of replicating traditional candidate approval polling. Although traditional polls are quite valuable sources of information, they are expensive and take time to gather. SM, on the other hand, can be gathered almost instantaneously, opening the possibility that SM could provide the ability to forecast ahead of polls. As with forecasting election outcomes, polling fore- and nowcasting can be built off features such as tweet volume and sentiment [129, 144, 171] or word choice [29].

In order to forecast both consumer confidence and presidential approval ratings, [144] gather tweets containing a small set of keywords and then measure tweet sentiment based on a previously available sentiment lexicon. Using the ratio of positive and negative sentiment on Twitter, they find a correlation both with Gallup polling on consumer confidence (released every three days) as well as with the Michigan Index of Consumer Sentiment (ICS) which is released monthly. In terms of forecasting, they explore the possibility of predicting the *next* month's ICS, finding a correlation of $r = 0.78$. This is worse than predicting by using previous ICS values ($r = 0.80$), but incorporating both Twitter and previous ICS features improves the correlation marginally ($r = 0.81$). Correlations are also reported by [144] in comparing Twitter sentiment about President Obama with Gallup job approval ratings, but [129] replicates these results on a slightly different range of dates and finds much poorer performance with the same method ($r = 0.22$ vs. $r = 0.73$). In order to improve their results, [129] decide not to filter only on tweets including the word "Obama", instead creating a political tweet classifier. Further, rather than using a keyword-based sentiment lexicon, as in [144], they create a supervised classifier which learns what vocabulary is associated with positive and negative sentiment based on the emoticons used in political tweets. Correlating the resulting sentiment ratio with Gallup polls, they report a final correlation of $r = 0.64$. Unfortunately, it is unclear whether this represents an advantage over simply using previous Gallup polls to forecast future poll results.

One difficulty in fore- and nowcasting public opinion using sentiment comes from the fact that not only are there a range of machine learning techniques which could be applied, but there are also any number of aggregation functions which could be used to represent sentiment. For instance, one might consider only the total number of tweets positively mentioning an entity. Alternatively, one could consider the ratio of positive to negative mentions. A wide variety of these functions are considered by [171], who use tweets to nowcast public opinion regarding five Portuguese politicians during the Portuguese bank bailout (2011-2014). They find multiple combinations of regression algorithms and sentiment functions which all converge on a similar level of performance, MAE 0.63%. In contrast to other work in the field, this represents a level of nowcasting performance which outperforms simply using previously published public opinion polls.

Given the conflicting results for both predicting election outcomes and polling data, does SM hold any power in predicting political outcomes? A combination of meta-analyses [78], literature reviews [76, 176], and editorials [75, 77, 82] have argued against the effectiveness of the predictions made above. Attempting to reproduce some of the above work, [78, 134] both fail to show that the proposed methods can consistently perform better than random chance. Additionally, research

into election prediction has exhibited a degree of confirmation bias and is subject to the effects of heavily biased populations [75]. Indeed, in the case of Twitter, the users who choose to engage in political discourse are quite rare and focusing on these users for prediction tasks introduces selection bias in all the analyses presented.

Despite an uncertain outlook, [78] remains hopeful that improvements can be made, and more powerful and useful models can be constructed for effective prediction in this domain. Indeed, in looking at the summary presented in Table 1, we see that the vast majority of prior work relies on fairly simple methods, ranging from standard linear models such as linear or logistic regression, to simple keyword matching. With recent advances in machine learning models, including ensembling and neural networks, there is a great deal to explore in applying these methods to SM data for election prediction. In addition to these methodological issues, more work needs to be done on actually *forecasting* these election events. That is, elections tend to be regularly scheduled, recurring events that can be planned for in terms of forming a predictive task. Launching real-time studies such as this would aid in researchers getting a more holistic picture of the state-of-the-art without overfitting a model to a validation set in a post-hoc analysis, as pointed out by [77]. Overall, this area of research has seen a great deal of investment and the limitations of past studies will likely help inform further research in this area, especially given recent interest in political polarization and its effect on SM interactions.

Table 1. Summary of Studies on Political Science

Article	**Topic**	**Data Source**	**Data Size**	**Features**	**Task**	**Success Rate**
Marchetti [129]	Approval rating	T	476M tweets	Semantic	Regression	$r = 0.71$ (approval)
O'Conner [144]	Approval rating	T	1B tweets	Semantic	Regression	$r = 0.81$
Saleiro [171]	Public Opinion	T	239K tweets	Semantic	Regression	MAE 0.63%
Tumasjan [191]	Election prediction	T	104K tweets	Semantic, Volume	Regression	MAE 1.65%
Skoric [181]	Election prediction	T	110K tweets	Volume	Regression	MAE 5.23%
Sang [172]	Election prediction	T	28K tweets	Volume, Non-SM	Regression	29% worse than polls
Cameron [27]	Election prediction	T, F	Not specified	Volume, Non-SM	Regression	Acc. 94.6%
Dokoohaki [61]	Election prediction	T	130K users	Social	Regression	$r = 0.65$ (Swedish), 0.79 (EU)
Pimenta [156]	Election prediction	T, F, Y, O	Not specified	Social, Volume	Regression	MAE 5.33% (T), 1.64% (F), 10.42% (Y), 1.67% (O)
Wang [200]	Election prediction	O	27K posts	Sentiment, Volume	Regression	MAE 4.0%
Khatua [102]	Election prediction	T	0.6M tweets	Volume	Regression	MAE 4.5%
Volkova [196]	Political party affiliation	T	1K users	N-gram, Social	Classification	Acc. 99.9%
Ceron [29]	Election prediction, approval rating	T, O	430K tweets	Semantic	Regression	MAE 2.4% (election), MAE 8-10% (approval)

T = Twitter, F = Facebook, Y = Youtube, FR = Flickr, O = Blogs, other

Stocks, Marketing, and Sales

Stocks

Of all SM forecasting tasks related to economics, predicting fluctuations in the stock market has been the most studied by far. Early work focused largely on predicting whether aggregate stock measures such as the Dow Jones Industrial Average (DJIA) would rise or fall on the next day, but forecasting can also involve making more detailed predictions, e.g., forecasting market returns or making predictions for individual stocks. The task is well studied outside of social media with a general consensus that forecasting is very difficult [14, 123, 127]. As in the case of elections, most research has focused on using general SM sentiment to make forecasts [22, 145, 147, 163, 223] although some papers investigate more nuanced models that learn the relationship between how individuals talk and market returns [35, 126].

The simplest task for stock market prediction is deciding whether the following day will see a rise or fall in stock prices. Comparison between studies is complicated by the fact that stock market volatility, and thereby the difficulty of prediction, may vary over time periods. High accuracy on this task was reported by [22], using sentiment analysis to achieve an accuracy of 87%. They find that measures of "calm" on Twitter along with DJIA numbers from the previous three days provide the best up/down predictions. Further adding the emotion "happy" reduces rise/fall accuracy to 80% but does reduce error in terms of forecasting absolute DJIA values. Importantly, they find that positive/negative sentiment analysis through the popular OpinionFinder tool [207] leads to no improvement over just using previous DJIA values. Their results are replicated by [130] and who forecast up/down movement for the French stock market with 80% accuracy. Removing sentiment, [128] use Tweets to forecast S&P500 movements with much lower accuracy (68%). Building on the techniques of [22], [163] uses an alternative sentiment analysis technique which is explicitly trained to learn what words on Twitter signal positive and negative sentiment. This does better than the OpinionFinder results reported by [22], achieving a rise/fall accuracy of 90.8% on the DJIA as compared to 60% when using historical data alone. A similar F1 score of 0.85 was reported by [145] when using sentiment from financial tweets rather than Twitter as a whole, while [35] reported slightly worse rise/fall accuracy (81%) in forecasting the Chinese stock market. Sentiment alone is not enough for good stock market forecasting. In particular, [157] make use of a dictionary-based sentiment method similar to the one used by [22]. Although their methodology does not appear profoundly different, their up/down classification performance is much worse with an accuracy of 64% on the DJIA, barely above the performance based on historical DJIA data reported by [22].

Stock market forecasts can also be made in terms of predicting the actual value of a stock or index rather than simply whether it will rise or fall. Evaluating both tasks, [163] found that the best rise/fall accuracy does not lead to the best accuracy in forecasting stock values. In particular, while the emotion "calm" works well in predicting rise/fall, the addition of the emotion "happy" both *reduces* rise/fall accuracy while *increasing* more fine-grained prediction. The choice of particular emotions for any analysis is emphasized by [223] who find that most emotions are poor predictors of future stock values. In their case, they find that *both* positive and negative emotions tend to lead to a decrease in stock prices, perhaps linked to the effectiveness of "calm" for [163]. Negative results are provided by [147] who investigate a variety of sentiment techniques to forecast stock values for nine US tech companies. No technique provides consistent improvements beyond a historical baseline, although they find Twitter is somewhat predictive of future trade volume and volatility. By also analyzing performance across a variety of tech companies, as well as composite indices, [163] similarly found that no method is predictive across all stocks.

More recent work has likewise found somewhat mixed results. Poor performance is reported by [224] who make use of a more complex, non-Gaussian statistical model. Their model forecasts the daily % increase or decrease in the DJIA and report a 33% root mean square error, meaning that average errors are potentially so large that either their methodology or report thereof is flawed. Building off of the mood analyses of [22], [119] introduce a variation which they use to

predict actual returns rather than up/down classification. Compared to the relative success of [163], the emotions used by [119] achieve much poorer results with the mood "sad" providing the best correlation results but with an r^2 of only 0.40. Lastly, [227] explore the possibility that noise in SM posts could be reduced by explicitly modeling stakeholders who affect a stock's price. To do this they draw from theoretical bases in economics and linguistics to first identify the features indicative of stakeholders (vs. other users) and then to make predictions based only off of stakeholder sentiment with $r^2 = 0.59$.

Taken altogether, work on stock market prediction is largely mixed. While the mood-based analyses pioneered by [22] have largely proven valuable, slight deviations away from their methodology have seen much less success indicating that the method itself may be unreliable. Further, while there has been a great deal of success in forecasting up/down movements in the stock market, the ability to gauge how large those daily shifts will be is a much more difficult task and has correspondingly seen less success. Another concern comes from the fact that almost all work in the area has been built off of [22]. Whether such a method represents the best which can be achieved from SM data is quite unclear. In particular, there is little evidence to suggest that SM data is more predictive of stock markets than other readily available predictors.

Marketing

An alternative use for SM forecasting is in the domain of marketing [94]. Although there is a great deal of work predicting what kinds of topics and products might go viral [8, 115, 202], we focus instead on a small sample of work which has been done in forecasting real-world outcomes. Very early work on blog posts demonstrated that blog posts about books showed little predictive power in determining whether Amazon sales would increase or decrease on the following day, but were useful in forecasting future sales spikes [87]. The authors speculate that this is because sales spikes are caused by outside events which are also captured through social media. Work by [37] shows how modeling latent user properties such as personality traits like openness and neuroticism can help companies create targeted advertisements. The authors created a Twitter account which posted travel information and recommendations relevant to users who post about their travel plans. They demonstrate that aiming advertisements at users with particular personality traits improves click- and follow-rates by 66% and 87% respectively, representing a large increase in value for companies.

SM has also been used to study the ability of online projects to successfully crowdfund their projects through websites like Kickstarter [121, 124]. Proposed projects create a page on Kickstarter asking users to donate generally small sums to fund the project. Users are enticed with rewards based on the level of their donation such as early access to the proposed item(s). Projects have a set fundraising goal as well as a project deadline. User donations are only made if the sum of all donations is higher than the goal by the time the deadline passes. If funding reaches the goal by this deadline the project is said to have been successfully funded, while if the deadline passes without the necessary amount donated the project is said to have failed and no money is given. In theory, SM should be predictive of crowdfunding success since projects are expected to succeed based on users sharing information about the project through SM. Work by [124] predicts whether a project will eventually succeed by making use of features relevant to the project itself (e.g., the fundraising goal), as well as social activity features (e.g., number of tweets related to the project), and social graph measures (e.g., average number of followers for project promoters). Using all of these features for only the first 5% of the project duration [124] achieved an accuracy of 76% in predicting whether the project will be successful. Similarly, success was shown by [121] even when just using SM information from the first three days of the project, achieving an AUC of 0.90, reflecting very high classification performance.

Movie Ticket Sales

Boosting movie ticket sales is an important task for marketing firms, and this has been studied specifically when marketing on social media platforms like Twitter. Indeed, the success of the 2016 film "Deadpool", having broken the record for the highest-grossing R-rated film of all time, is often attributed to its social media marketing strategy [150].

Previous research linking SM to movie sales has demonstrated somewhat less predictive power than might have been anticipated. When [136] correlated box office sales for particular movies with SM information they found only moderate correlations ($r^2 = 0.29$) when using positive sentiment on blogs along with volume of blog posts. This represents a 12% increase over using the volume of blog posts, but still is far from impressive. Better results are reported for a volume-based analyses by [3]. They predict daily box office revenue for movies as well as sales rank for music albums achieving $r^2 = 0.74$ and 0.92, respectively. Baseline features such as movie budget, genre, or number of theaters, which may hold greater predictive value, were not provided by [136]. Further work on movie sales was done by [186] who monitored official Facebook fan pages for 50 different movies. They achieve an r^2 of 0.88 in forecasting total box office revenue when incorporating social network features, essentially modeling the influence of each movie's fan page, a significant improvement over using just the number of theaters showing each film ($r^2 = 0.68$).

Much more positive results are reported by [9] who correlate Twitter volume with opening weekend box office sales achieving $r^2 = 0.93$ with SM data alone and $r^2 = 0.97$ when incorporating SM with the number of theaters showing a film. Again, what is missing from their analysis is any systematic comparison of SM features with the kinds of non-SM features that would be used in any serious forecasting attempt. This limits our ability to determine the real predictive value of SM over-and-above baseline features. Further, a lack of systematic comparisons between various SM platforms makes it difficult to compare studies against one another or to know which platforms researchers should focus on in the future.

The relative scarcity of publications in this area of social media data analysis suggests that this is a rather difficult area of investigation. While many of the studies surveyed here present positive results, it is worth noting that many of these studies also opt to report correlations between model predictions and some ground-truth signal. These measures may obscure more nuanced model behavior as in more controlled machine learning experiments that use more sophisticated measures such as *average precision* or ROC area-under-the-curve metrics. At a more qualitative level, the number of studies focusing on sentiment as a key indicator of stock or sales performance is striking. This presents many limitations and difficulties given that sentiment detection is still a somewhat open research area, and much past work casts sentiment as a crude distinction between positive or negative polarities [83].

Given the overall dearth of work in these areas, it is difficult to assess which of these areas may hold more promise over the others in terms of future research. In all cases, studies purporting to predict economic response variables by incorporating social media benefit from additional features outside social media, such as other economic indicators. This is particularly difficult when utilizing Twitter data, where the vast majority of tweets that can be collected will not mention a product of interest or the stock market. These challenges, along with a lack of deep understanding of how users interact online with respect to economic phenomena, will likely make it necessary to incorporate data outside social media in order to build accurate models in this application area.

Public Health

Significant effort has been made in utilizing SM and other Internet data for the purpose of monitoring, predicting, and improving public health. Research on using SM for public health addresses a wide range of phenomena, including monitoring and forecasting disease outbreaks, identifying individuals in need of mental health services, and identifying specific adverse drug effects before they were discovered by the U.S. Food and Drug Administration (FDA) [41, 71, 139]. An overview of how Internet data in general can be used in the public health domain is given

Table 2. Summary of Studies on Economics

Article	Topic	Data Source	Data Size	Features	Task	Success Rate
Chen [37]	Advertising	T	5.9K users	Semantic	Regression	66% gain (click rate), 87% gain (follow rate)
Li [121]	Crowdfunding success rate	T, F, K	106K tweets	Metadata, N-gram, Social	Regression	AUC 0.90
Lu [124]	Crowdfunding success rate	T, K	Not specified	Metadata	Classification	Acc. 76%
Gruhl [87]	Product sales rank	O	300K blogs	N-gram	Classification	Acc. 63%
Bollen [22]	Stock market	T	9.8M tweets	Semantic	Classification	Acc. 80%
Chen [35]	Stock market	SW	256K tweets	N-gram	Classification	Acc. 81%
Makrehchi [126]	Stock market	T	2M tweets	Semantic	Classification	20% gain (returns)
Oh [145]	Stock market	O	208K blogs	Metadata, Semantic, Non-SM	Classification	F1 0.85
Mao [128]	Stock market	T	Not specified	Volume	Classification	Acc. 68% (S&P500)
Porshnev [157]	Stock market	T	755M tweets	Semantic	Classification	Acc. 64% (DJIA), 62% (S&P500)
Martin [130]	Stock market	T	173K tweets	Semantic	Classification	Acc. 80%
Oliveira [147]	Stock market	T	Not specified	Semantic	Regression	$r^2 = 0.20$
Rao [163]	Stock market	T	4M tweets	Semantic, Non-SM	Regression	$r^2 = 0.95$ (DJIA), $r^2 = 0.68$ (NASDAQ)
Zimbra [227]	Stock market	O	64K posts	Semantic, Non-SM	Regression	$r^2 = 0.59$
Li [119]	Stock market	T	Not specified	Semantic	Regression	$r = 0.63$ (sad), 0.49 (anger)
Zhao [224]	Stock market	T	Not specified	Semantic, Volume	Regression	RMSE = 33.0% (DJIA)
Mishne [136]	Movie sales	O	Not specified	Semantic, Volume, Non-SM	Regression	$r^2 = 0.29$
Asur [9]	Movie sales	T	2.9M tweets	Semantic, Volume, Non-SM	Regression	$r^2 = 0.97$
Tang [186]	Movie sales	F	Not specified	Social, Non-SM	Regression	$r^2 = 0.88$
Abel [3]	Movie & album sales	O	100M posts	Volume, Non-SM	Regression	$r^2 = 0.74$ (movies), 0.92 (albums)

T = Twitter, F = Facebook, SW = Sina Weibo, K = Kickstarter, O = Blogs, other

by [90], and [17] gives a chronology of developments in utilizing SM data. Early work, as surveyed by [169], identified the potential utility of incorporating SM into public health-related tasks. More recent comprehensive reviews confirm this potential while noting the lack of actual systems taking advantage of SM [34, 85, 193].

Influenza

Success in predicting epidemiological outbreaks was reported by [45, 47, 90, 117] to varying degrees. A canonical example of sentiment and time series analysis in Twitter over the 2008-2009 influenza season in the United States was provided by [46]. The authors report a high correlation between queries for curated vocabularies in Twitter data and influenza-like illness prevalence in the United States. Work by [90] shows that an outbreak of dengue fever on Madeira Island (a Portuguese territory) was tracked in real-time using online biosurveillance techniques. Work by [117] utilizes an agent-based model [23] but reports ambiguous predictive power with data collected from a purpose-built application. When used in conjunction with and validated by traditional data sources via the Center for Disease Control (data specifically from the Outpatient Influenza-like Illness Surveillance Network), Twitter data can reduce forecasting error by 17-30% [151] reports. Specifically, [108] combines part-of-speech tagged, stemmed, and Amazon Mechanical Turk labeled Twitter data with external sources like Google Flu Trends to gain this increased forecasting resolution. In related work, [51] find that models augmented with Twitter n-gram and LIWC features are more accurate in predicting 20 county-level health-related statistics.

Although not expressly an SM data source, Google Flu Trends (GFT), released in 2008, has been the topic of much discussion in the literature, and is often used as a basis for comparison in Internet data-based biosurveillance models [193]. Work by [45] gives an early assessment of GFT's ability to predict influenza outbreaks by monitoring the search prevalence for influenza-like illness symptoms, showing promising results and supporting early excitement in epidemiological research. However, [148] subsequently reevaluate these results, and show that because GFT leans so heavily on correlative measures assumed to be good predictors, the models developed ultimately did not anticipate the 2009 H1N1 pandemic and severely overestimated both the 2011-2012 and 2012-2013 flu seasons. In fact, [111] shows that GFT overestimated the number of anticipated cases in the 2011-2012 season by more than 50%. This could be explained by a shift in public attention to influenza following the 2009 pandemic. As [111] note, GFT makes the assumption that online behavior is determined by outside events (e.g. illness) but does not take into account the way online platforms shape the way users search. For instance, if a search platform suggests to a user that a query related to "fever" or "cough" might be flu-related, this may influence the user to continue searching for influenza-related information. These additional queries could bias GFT to believe the flu is more common or severe than it actually is.

Some doubt regarding studies which correlate SM with influenza-like illnesses is cast by the work of [19]. They expand upon previous work on Twitter demonstrating that high correlations may in fact be the result of questionable methodology. They replicate three studies, demonstrating that the similar, or even better, performance can be achieved using irrelevant or falsely generated data. This suggests that the mathematical models used in previous studies may be too powerful, overfitting the small amount of real-world influenza data which should lead to difficulty in generalizing the model to new data (as seen with GFT). This position is strengthened by the fact that the same models generalize very poorly when trained on one spatial region and tested on another. That is to say, a model trained on data from the US Northeast will likely perform very poorly in forecasting influenza-like illness on the West coast.

Echoing the positions taken by elections researchers [78, 134, 176], without a better understanding of predictors in SM data, developing accurate models of external events based on SM features will continue to be quite difficult. There are any number of powerful models which can be used to model illness in the real world, but it is unclear how well any of these methods might generalize across space or time [19]. That being said, research on disease detection and forecasting continues both for influenza-like illnesses and other diseases. For instance, [228] apply advances in deep learning to the task of detecting infectious intestinal diseases such as norovirus and food poisoning.

Mental Health

Infectious diseases are not the only health issue with relevance to SM. Researchers have also begun to use SM to identify or predict various mental health issues including addictive SM usage [180], anorexia recovery [32], addiction recovery [140], distress [114], suicidal ideation [57], suicide rates [209], and post-partum depression [52, 54], as well as depression more generally [55, 165, 188].

A chief challenge in the area of mental health disorders is getting help to individuals in need. Screening for mental health problems is expensive and many disorders may make individuals less likely to seek out professional help. Social media promises a cheap, and possibly immediate, method of identifying individuals who may benefit from outreach. While current work suggests SM may hold great promise, there are also a number of limitations involved, including a lack of systematic reviews, differing methodologies, and difficulty in creating a ground-truth for model comparison.

Of all mental health disorders, depression has received the greatest attention from SM researchers. In the works of [52–55] postpartum and general depression was studied among Twitter users taking advantage of behavioral features (e.g., volume of tweets, number of replies) as well as linguistic features (e.g., positive and negative sentiment, use of pronouns). In all three studies, ground-truth for depression was measured by having each user fill out a survey on depressive symptoms. The reported studies were able to achieve accuracy rates of 72.4% [55], 74.6% [53], and 80.5% [52] in identifying depressed users. This work finds that depressed users on SM can be characterized by decreased social activity as well as increased negative sentiment and use of personal pronouns. In order to validate their findings, [54] conducted interviews with the 165 subjects of their study. The authors found that reduced social interactivity alone explained up to 50% of the variance in collected data. Similar results are also found for Japanese Twitter users, possibly reducing concerns that previous work on English-speaking users would not be generalizable [188].

Social media data has also been used to make predictions about recovery. [32] use survival analysis to examine Tumblr users who self-identify as anorexic. Even using very simplistic behavioral and linguistic features, they are able to predict recovery higher than chance. They identify specific features that predict recovery and compare these against features suggested by previous literature. Work by [140] applies a similar technique to Twitter users attempting to overcome nicotine addiction. Their features and model are also quite simplistic but are still able to show clear, statistically-significant differences between relapsers and those who successfully quit smoking. While exploratory, their work does suggest that simple features tied to the existing domain-specific literature may contain the signal necessary for proper classification. This is supported by quantitative results presented by [57] who attempt to predict if posters on mental health forums on Reddit will show signs of suicidal ideation. They report an accuracy of 80% in predicting whether users will begin posting in the next few months to the subreddit r/SuicideWatch, a forum for users thinking about committing suicide. Compare this to the work of [26] who classify posts related to suicide based on their intent. For example, some mentions of suicide indicate suicidal intent while others may be a report of suicide, condolence, or a flippant reference to the act. With seven classes, they achieve an F1 score of 0.69 which indicates relatively good classification on this task. A similar study focusing on nowcasting was able to use Reddit posts to identify individuals who were distressed [114]. They report an accuracy of 54.5% versus a baseline of 30.5% when classifying four ways based on the level of distress.

The value of more complex models and features is demonstrated by more recent work [31, 180]. The work of [180] makes use of behavioral features related to Instagram and Facebook usage in order to detect users with a social network mental disorder, mental disorders which manifest themselves in terms of "excessive use, depression, social withdrawal, and a range of negative repercussions" [180]. They achieve classification accuracy as high as 92.6% through the use of a more complex model which better takes into account changes in behavior over time. By comparing a number of machine learning techniques, [180] are able to show the importance of choosing an appropriate model, suggesting that previous results may be particularly limited by

the less advanced techniques often employed by social scientists. While much of the work on mental health fore- and nowcasting relies heavily on hand-curated lexical features, [31] provide an alternative by combining a statistical technique known as topic modeling with insights from clinical annotators. By identifying the topics in SM posts related to eating disorders, they are able to classify the severity of a users eating disorder with high accuracy ($F1 = 0.81$).

Taken together, existing work on mental health disorder detection and prediction suggests that SM is a valid and useful tool. Classification performance ranges from mediocre to very good with the greatest success in areas where more advanced features and models have been used. While the greatest number of papers have been published on detecting depression, existing work also relies almost exclusively on very simple behavioral and linguistic features within a logistic regression framework. Given the strong performance of [180], there may be room for improvement within this domain if researchers are willing to apply more advanced techniques over larger quantities of data.

Much of the work on diagnosing mental health disorders from SM data makes use of techniques that can be applied elsewhere. A good example of this is [58] who makes use of topic modeling to understand users food choices based on the food items they post to Instagram. They make use of these food topics and user geolocation information in order to detect whether a particular region is a "food desert", an area with limited access to nutritional food items. They achieve an accuracy of 80% in this prediction task by combining both user posted information with publicly available socio-economic data from each region.

While applying more advanced statistical techniques and machine learning is a clear area for improvement, there remain a number of methodological difficulties that future work must address. Perhaps the greatest difficulty is in comparing against ground-truth data. Diagnosing mental health symptoms can only be done by trained professionals, which makes it difficult to know whether a particular SM user has a disorder or not. As a result, researchers typically focus their work on particular users who volunteer to fill out a survey measuring these symptoms. While this technique provides researchers with a ground-truth, it biases these studies, making it difficult to know if results can be applied to all SM users.

Adverse Drug Reactions

Information from SM has also been used to identify drug users suffering from adverse drug reactions (ADRs), negative side effects arising from pharmaceutical drugs taken as prescribed. Given that ADRs are typically reported on a case-by-case basis to physicians, the ability to monitor online disclosure of these reactions at a larger scale could greatly increase the ability of medical professionals to intervene and track these cases. Most studies in this area focus on data mined from online medical forums where individuals ask questions about their symptoms [71, 142, 178], although one study also evaluates ADRs using Twitter [18].

Detection of ADRs first requires the identification of pharmaceutical drug users. Researchers have made use of publicly available medical dictionaries in order to train classifiers, achieving reasonable accuracy. In the work of [18] a medical dictionary is used to identify drug-related tweets and they are able to identify drug users with a mean accuracy of 74% and AUC of 0.82. [178] achieve somewhat better results using online health forum data, achieving 87% precision and 80% recall.

In terms of identifying ADRs, [18] use a bag-of-words approach, classifying with a support vector machine (SVM) and achieve a mean accuracy of 74% and area under the curve (AUC) of 0.74, somewhat lower than their results for drug user identification. Somewhat lower results in ADR identification were also achieved by [178] with a precision of 85% but much lower recall at 56%. A classifier was trained by [142] using association rules based on keywords and part-of-speech tags. They also found the task difficult, reporting a precision of 70% and recall of 66%. Together, these studies indicate that identifying ADRs can be accomplished somewhat successfully using a wide range of features and classification techniques. After testing three different methods for extracting ADRs from both medical forum and Twitter data [214] found

best results in applying an initial filter to posts followed by a sequential model that is able to extract actual mentions of ADR-related terms from the text.

The previously mentioned studies show that SM can potentially be used to identify drug use and ADRs. Taking this research a step further [71] produced an unsupervised PMI-based classifier and used it to predict ADRs not labeled by the FDA. Their system models drug-symptom relations and classifies a symptom as an ADR if it appears more often in user comments than expected by chance. They tested their model on two case studies, cholesterol-lowering drugs and anti-depressants. Cholesterol-lowering was chosen because the class of statin drugs were relabeled by the FDA in 2011 to have cognitive impairment as a possible side effect. Among anti-depressants, Wellbutrin was relabeled in 2009 to include agitation as a possible side effect. Using their model [71] were able to correctly identify both of these relations using user comments before the FDA relabelings. Among all the ADRs in their data, they achieved a high precision of 93.2% with recall of 70.4%.

This work shows that identification of ADRs is possible using online comments, particularly through health forums. Degree of success is mixed, likely due to varied methodology and reliance on properly integrated medical terminology databases. Still, current work suffers from a number of difficulties which might be improved. First, feature selection has generally been rudimentary, using bag-of-words [18], dictionary-based keywords [178], and simple association rules [142]. The best performance, achieved by [71], makes use of a more complex grammatical parsing algorithm along with relational modeling of drugs, symptoms, and individuals. The success of future work likely hinges on incorporating more robust techniques from machine learning.

Table 3. Summary of Studies on Public Health

Article	Topic	Data Source	Data Size	Features	Task	Success Rate
Bian [18]	Adverse drug reactions	T	239 users	N-gram, Semantic, Non-SM	Classification	Acc. 74%
Feldman [71]	Adverse drug reactions	O	41K posts, 5.3K users	Semantic, Non-SM	Classification	F1 0.84 (statins) F1 0.78 (anti-depressants)
Nikfarjam [142]	Adverse drug reactions	O	6.8K posts	Semantic	Classification	F1 0.68
Segura [178]	Adverse drug reactions	O	400 posts	Semantic, Non-SM	Classification	F1 0.68
Yates [214]	Adverse drug reactions	T, O	400K forum posts, 2.8B tweets	N-gram, Semantic, Non-SM	Classification	Prec. 0.59 (O) Prec. 0.48 (T)
Corley et al [46]	Influenza	T, O	97.9M posts	Metadata, N-gram	Regression	$r = 0.63$
Lamb [108]	Influenza	T	3.8B tweets	N-gram, Semantic	Regression	$r = 0.80$
Paul [151]	Influenza	T	Not specified	N-gram, Semantic	Regression	25.3% improvement
Bodnar [19]	Influenza	T	239M tweets	N-gram	Regression	$r = 0.88$
Zou [228]	Intestinal disease	T	410M tweets	N-gram	Regression	$r = 0.73$ (Norovirus), 0.77 (Food poisoning)
Zhang [222]	Asthma	T	5.5M tweets	N-gram	Classification	Acc. 66.3%
Chancellor [32]	Mental health	TR	13K users, 68.3M posts	Metadata, Semantic	Regression	Concordance 0.658
De Choudhury [52]	Mental health	T	40K tweets	Semantic, Social	Classification	Acc. 80%
De Choudhury [55]	Mental health	T	2.1M tweets	Semantic, Social	Classification	Acc. 70%
De Choudhury [54]	Mental health	F, T	40K tweets, 0.6M posts (F)	Metadata, Semantic, Social	Regression	$r^2 = 0.48$
De Choudhury [57]	Mental health	R	63K posts, 35K users	Metadata, Semantic	Classification	Acc. 80%
Burnap [26]	Mental health	T	2K tweets	N-gram, Semantic	Classification	F1 0.69
Shuai [180]	Mental health	F, I	63K users (F), 2K users (I)	Metadata, Social, Behavior	Classification	Acc. 78% (I), Acc. 83% (F)
Tsugawa [188]	Mental health	T	209 users, 574K tweets	N-gram, Semantic, Social	Classification	Acc. 66%
Won [209]	Mental health	O	153M posts	N-gram, Non-SM	Regression	Acc. 79%
Chancellor [31]	Mental health	I	100K users	Semantic	Classification	F1 0.81
Lehrman [114]	Mental health	R	200 posts	N-gram, Sentiment	Classification	Acc. 54.5%, baseline 30.5%
Culotta [50]	Public health statistics	T	4.3M tweets	Metadata, Semantic, Non-SM	Regression	$r = 0.63$
De Choudhury [58]	Food Deserts	I	14M posts	Semantic, Spatial, Non-SM	Classification	Acc. 80%

T = Twitter, F = Facebook, SW = Sina Weibo, I = Instagram, TR = Tumblr, R = Reddit, O = Blogs, other

Threat Detection

Numerous attempts have been made to use Twitter data to detect rare or anomalous real-world events such as natural disasters, security events, and political uprisings. These types of events have garnered attention due to their implications for safety and security, and the spontaneity with which they arise. That is, unlike many of the other event types surveyed here, such as elections or the spread of influenza, these events do not occur regularly and do not have a limited set of outcomes (e.g., winning or losing an election). Rather, these events often constitute crises or disasters that an automated system should be able to detect in real-time as opposed to forecasting into the (distant) future.

Threat detection has built on more general work in event detection which aims to identify events from a stream of SM posts. For example, [199] focus on automatically identifying geographically localized events via Twitter streams using a combination of geofiltering and clustering techniques. Individual tweets are assigned to clusters based on 41 features computed from tweet text and geolocation metadata, and the resulting clusters are classified as belonging to either an "event" or "non-event" class. The authors report an *F1-score* of 0.857 using a pruned decision tree. An alternative system, *EvenTweet*, is proposed by [2]. They incorporate geolocation features as well as similarity between keywords over a particular time range to identify tweets corresponding to real-time events. Although they tested their model specifically on soccer matches, their framework is general enough to be potentially applicable to other domains. The literature on general event detection is quite large and spans any number of specific content domains [6, 7, 60, 133, 149, 225], therefore we focus specifically on event detection in the context of security events.

Cybersecurity Events

Cybersecurity, otherwise known as IT or computer security, is an increasingly important area of interest for the protection of national, corporate, and organizational interests. A review of issues and state-of-the-art techniques in cybersecurity is well outside the scope of this literature review. Interested readers may find more information in recent literature reviews such as [73, 125]. Cybersecurity research involving SM can be thought of in terms of two lines of work: 1) can SM be used to detect cybersecurity events on SM systems? 2) can SM be used to detect cybersecurity events affecting other systems?

Addressing the first question, [97] propose a method for detecting "identity cloning accounts" on Facebook. The goal of these attacks is to enter into actual users' friendship circles to access privileged information. Although their detection algorithm appears feasible, they are limited in evaluation due to the fact that they cannot collect information about actual attacks nor can they simulate an attack on Facebook themselves. In the work of [210] they sought to circumvent these issues by having actual Facebook users participate in an experiment, browsing their own Facebook pages as well as browsing a stranger's. Given that intruders are likely to show different click behavior than legitimate users, they propose a detection scheme based on Smooth Support Vector Machines (SSVM). After two minutes they are able to identify intruders with an accuracy of 81.9% and by nine minutes 92.9%. While these results are impressive, savvy intruders might be able to circumvent the system by modifying their own behavior.

In terms of using SM to detect cybersecurity events happening outside of SM, current work has achieved less success. [166] explore the ability of Twitter to detect data breaches, account hijackings, and distributed denial of service (DDoS) attacks. The biggest challenge for such an attempt is the problem of data sparsity. While the volume of total tweets is quite large, tweets relevant to cybersecurity events are rare and tweets related to *current* cybersecurity events are even less common. In spite of these challenges, [166] are able to detect these types of events to some degree. They present their results in terms of area under the curve for precision-recall plots, achieving scores of 0.716, 0.676, and 0.459 for account hijacking, data breaches, and DDoS attacks respectively.

Research on detecting cybersecurity events mirrors work on SM seen elsewhere. The most impressive results come when SM behaviors are used to detect irregular activity representative of an account hijacking as in [210]. Behavioral markers of the type they utilize appear to be quite powerful in prediction, but wide-scale analysis of such systems cannot be done by outside researchers. This limits, unfortunately, the usefulness of such measures. On the other hand, it is possible to detect when SM users are discussing cybersecurity events [166]. Work on this topic is in its infancy, relying on topic and sentiment analysis techniques which are still under active development.

Protests, Civil Unrest, and Crime

While cybersecurity events typically occur online, SM can also be used to detect and predict offline events. Of particular importance for threat detection are the prediction of events such as crimes, protests, and other types of civil unrest. Because mass use of SM has emerged only recently, research focuses largely on a small number of geographical regions where civil unrest has occurred in the past decade, particularly the Middle East [20, 101, 184, 203] and Latin America [84, 141, 162, 212], although some work exists on protests within the United States [56]. Whereas cybersecurity events are detected in real-time, i.e., "nowcasting", protests and civil unrest require coordination among individuals and research has investigated the ability not just to detect these events but also to forecast when they will occur.

Work on the Middle East has focused largely on the so-called "Arab Spring" in early 2011 [20, 184, 203], although [101] examines the 2013 military coup in Egypt. The work of [20] examines the ability of information mined from Twitter to detect and predict Arab Spring protests in Egypt within a timespan of 3 days. They use tweet content and "follower" relations on Twitter in order to predict events taken from GDELT [113], a publicly available database of political events around the world. They achieve similar results both for detection and predicting within a 3 day timespan with accuracies of 87.1% and 87.0%, respectively [20]. They report that social relations alone achieve much better classification than tweet text content. The work of [203] examines a similar Twitter dataset with an alternative probabilistic graphical model. Although they do not report quantitative accuracy scores for their model, they come to a similar conclusion that including additional information beyond just tweet content increases model performance. That being said, even simple statistical relations can be used to predict protests. The work of [184] demonstrates that coordination on SM (e.g., the adoption of a small set of frequently used hashtags) is predictive of protest volumes on the following day, not just in Egypt but throughout all Arab countries majorly affected by the Arab Spring.

Little work has been done to extend the research on civil unrest detection to more developed, Western countries. This is perhaps due to the relatively decreased civil unrest within these nations. With the rise of the "Black Lives Matter" movement and worries about police shootings in the USA, there has been a single study attempting to extend this work [56]. This research is mainly exploratory, investigating SM variables that might have been used to predict protest volumes across geographic regions. In terms of accuracy, their system captures 81% of protest volume within 20% of the true value, which is perhaps impressive given their use of a simple Poisson regression model using limited features.

Separate research has been conducted using ground-truth data about protests, riots, and civil unrest in Latin America focusing on a timespan between 2013 and 2014 when major protests took place in many countries. Work by [212] and [84] is similar to previously examined work on the Middle East in that they both attempt to retrospectively build systems capable of detecting unrest. Working with data from Tumblr [212] demonstrated impressive precision in event detection (95.6%) with an average lead time of 4.8 days. The work of [84] looks at Twitter data attempting to model cascading social participation, the idea that individuals are more likely to join an offline protest if they are exposed to a critical mass of online support. Their work is somewhat less successful, achieving an F1 score of approximately 80% on their Brazilian data, but only 55% on their Venezuelan data. It should be noted, however, that both of these models are

trained on data from time periods of heavy unrest, and it is unclear whether these models would have predictive power in an implemented event detection system.

A separate line of research on Latin American protests is well worth discussing in terms of clarifying how well a fully implemented system might be capable of detecting unrest events. The EMBERS system, as described by [162], is an automated event detection system which has been in place making forecasts since November 2012. EMBERS makes use of a wide variety of publicly available data including Tweets, RSS news and blog feeds, meteorological data, and Google Flu Trends among many others. While most research papers focus on a single model of event detection, EMBERS makes its forecasts by combining five separate models, only two of which has been fully investigated in separate work [105, 141]. They evaluate their models' forecasts based on a "quality score" which equally weights predictions in terms of how well they match the date of the event, the location, the event type, and the population which will be involved in the event. This results in a score from 0 to 4, with higher values indicating better predictions. An average quality score of 3.11 is reported by [162], with recall of 82%, precision of 69%, and an average lead time of 8.88 days. Putting this into context, the results of a single model achieves a classification F1-score ranging from 0.68 to 0.95 depending on country [105]. These impressive results, not just in detecting events but in detecting specific properties of those events, demonstrate just how powerful the lines of research previously discussed might be if put into practice.

While SM may be predictive of large-scale organized protests and civil unrest, one may wonder whether the same would be true for smaller-scale criminal acts which do not have the same organizational requirements. Published work on crime forecasting is largely limited by the (in)accessibility of large-scale databases of detailed criminal records. Because of this, published research has largely focused on a single city, Chicago, which has made some of its records available for research purposes [39, 80]. A wide variety of crime types are examined by [80]. The simplest method for predicting areas where crimes are likely to take place involves making use of historical crime data. This can be augmented by SM posts, which [80] analyze using a topic model, linking the learned topics to the likelihood of various crimes. Of the 25 crime types examined, 19 see a forecasting improvement by incorporating SM data. Using the same data, [39] choose to combine historical crime data with a sentiment analysis of Twitter posts as well as with current weather conditions. They report classification accuracy somewhat lower than found in [80] and also find that the addition of weather and SM sentiment only improves classification marginally. Further, because they do not report separate results for SM data or weather data alone, it is unclear whether this marginal improvement is in fact the result of adding SM data at all. Based on these findings, it is unclear what the best approach to making use of SM data for crime forecasting might be. While there may be some value for police organizations, it is unclear how great that value might be.

Natural Disasters

Natural disasters represent an additional type of threat which can be detected through SM. One of the first demonstrations of SM's ability to detect events in real time came in the form of earthquake detection [104]. Although earthquakes cannot be *forecasted* through SM, a number of articles have demonstrated that Twitter users represent a sensor network that can potentially outperform standard seismology systems [49, 66, 168]. In these special cases, earthquake-related tweets can functionally predict when standard seismology equipment will detect an earthquake a few moments later. Not only can Twitter beat traditional sensors, but they can also distribute information about a quake to potentially affected individuals much faster than governmental agencies [168]. Additionally, SM data can be used to detect earthquakes where high resolution equipment may be sparse [69]. Other types of natural disasters are explored by [135] who particularly focus on identifying specific locations most in need of aid. Applying their method to earthquakes, floods, and tornados, they are able to identify streets and places of interest most hard hit by disasters.

A small set of studies also exist on using SM to now- and forecast weather more generally.

Concerns regarding climate as a national security issue have made these additional studies increasingly relevant in the area of threat detection [143]. While this is an underexplored area, there is some evidence that SM can be used to both detect and forecast weather. For instance, the authors of [109] use tweets from 5 UK cities to nowcast amounts of rain. Even with a relatively simple model based on weather-related keywords they are able to nowcast daily rainfall with an RMSE of 2.6mm, a 40% error reduction versus their baseline model. Levels of air quality can also be inferred through SM data [36, 120, 132]. Looking at air pollution levels in Chinese cities, [132] first build a spatial model which does not take advantage of SM. This allows the model to take advantage of the fact that pollution levels across cities are correlated based on their real-world properties. Using this as a baseline, they expand to include information from Sina Weibo posts finding a 13% reduction in prediction error. The problem of air pollution detection can also be treated as a detection task where there is either a smog event in a city or not on any given day. Taking this approach, [36] combine traditional physical sensors with SM check-in and textual data in order to estimate the population mobility and traffic conditions which might lead to smog events. Their neural network-based architecture is able to classify smog events with high accuracy using only physical sensor data, but sees an improvement in performance when adding SM data as well, indicating that even when traditional sensors are in place SM data can still add additional value. While the previous two works focus on textual data from SM, [120] make use of images posted on SM in order to monitor air pollution. While their method works very well on a small set of high quality images ($r = 0.89$), on a larger set of noisy images taken from SM the method is much less successful ($r = 0.41$).

Threat detection is an area of great promise with at least one detection system currently operational and demonstrating great effectiveness. Positive results through retrospective analysis utilizing a variety of modeling techniques also show the possibility of real-world usage. At the same time, it should be made clear that extremely successful results in retrospective studies, as in [113, 212], are unlikely to translate into similar results for actual forecasting. This is because many of these retrospective analyses formulate the supervised learning problem as only having to decide between a small set of classes, in some cases just two (e.g., event or non-event). This greatly simplifies the learning problem, but real-world systems such as EMBERS [162] need to make predictions not just of time and event, but also of event type, population involved, and location. These added details greatly increase the difficulty of the problem and likely require incorporating information from outside of SM, such as weather or financial news [162].

Table 4. Summary of Studies on Threat Detection

Article	Topic	Data Source	Data Size	Features	Task	Success Rate
Boecking [21]	Civil unrest	T	1.3M tweets	Metadata, N-gram, Semantic	Classification	Acc. 87%
Boecking [20]	Civil unrest	T	1.3M tweets	Metadata, Semantic, Social	Classification	Acc. 87%
De Choudhury [56]	Civil unrest	T	29M tweets	Semantic	Regression	$r^2 = 0.42$
Kallus [101]	Civil unrest	T, O	Not specified	Semantic	Classification	AUC 0.92
Ramakrishnan [162]	Civil unrest	T, O	3B messages	Semantic, Social, Spatial, Non-SM	Classification	Rec. 0.82
Korkmaz [105]	Civil unrest	T	500M tweets	N-gram, Non-SM	Classification	F1 0.95 (Brazil), 0.88 (Mexico), 0.70 (Argentina)
Gerber [80]	Crime	T	1.5M tweets	Semantic, Spatial, Non-SM	Classification	AUC 0.72
Chen [39]	Crime	T	1.0M tweets	Semantic, Non-SM	Classification	AUC 0.67 (w/ SM), 0.66 (w/o SM)
Wu [210]	Cybersecurity	F	112 users	Behavior	Classification	Acc. 90%
Ritter [166]	Cybersecurity	T	14.6M tweets	Semantic	Classification	AUC 0.72 (hack), 0.46 (DDoS), 0.68 (breach)
Avvenuti [10]	Earthquakes	T	Not specified	N-gram	Classification	F1 1.00 (magnitude $\geq$4.5)
Sakaki [168]	Earthquakes	T	Not specified	N-gram, Semantic, Spatial	Clustering	MAE 3 deg. (lat./long.)
Middleton [135]	Natural disasters	T	Not specified	N-gram, Spatial, Non-SM	Classification	F1 0.77 (Hurricane), 0.53 (Tornado)
Mei [132]	Weather	SW	Not specified	N-gram, Spatial	Regression	13% gain w/ SM
Chen [36]	Weather	SW	Not specified	Semantic, Spatial, Non-SM	Classification	AUC 0.976 (current smog), 0.956 (current no smog)
Li [120]	Weather	O	8.7K images	Visual	Regression	$r = 0.41$
Lampos [109]	Weather	T	8.5M tweets	N-gram	Regression	RMSE 2.6mm

T = Twitter, F = Facebook, SW = Sina Weibo, O = Blogs, other

User Characteristics

While many of the areas of study previously investigated are more closely linked to predicting events in the real world, many of these tasks rely on or might be improved by accurately inferring various user characteristics. For instance, political preferences vary across sub-populations with not all populations equally represented in SM data [38]. If researchers could leverage inferred demographic information about users, they might be able to de-bias their models or otherwise take advantage of known demographic relationships, such as the weighting scheme employed in [172].

Because of the importance of predicting user characteristics, this has been an active area of research on its own, addressing a number of different prediction tasks. Potential applications have driven interest in this area of research given the possibility of major breakthroughs. For instance, security analysts might use real-world demographics to understand membership in online groups

of interest, and marketers could better take advantage of "customer segmentation" where user locations and demographics correlate highly with purchasing certain products.

Work in each of these areas has utilized text produced by users within social media platforms [33, 40], as well as images [215], and user metadata [89, 177, 195] to build models for location and demographic prediction. While the studies across these areas no doubt present challenges for future work, the positive results reported indicate sustained progress in incorporating a variety of features to build ore accurate and predictive models.

Geolocation

Prediction of user location has been an active area of research, especially given that many social networks may contain very little geolocation metadata. The work of [40], for example, notes that fewer than 1% of tweets in their study contain geolocation tags. Of U.S. Facebook users in 2010, only 4% had entered their home address in a way that could be converted to latitude and longitude coordinates [40]. Because most users are not sharing their location information, any attempt at geolocation will necessarily be biased by the characteristics of those whose data is available.

Geolocation research can be roughly divided into two separate prediction tasks. On the one hand, location can be thought of as a static property of each user corresponding to their home address. On the other, location can be thought of as constantly shifting, such that researchers might attempt to predict where an individual is likely to be located in the future. It should be noted that we focus predominantly on mean error distances rather than medians. While many studies achieve impressive results in terms of *median* errors, mean distances are often quite larger because users who are not located with high precision are often very poorly located. In terms of applicability, readers should keep in mind that an average error distance of 300 miles may seem quite large, but may still reflect a model that classifies half of users accurately within only a handful of miles.

Multiple methods can be used to predict a user's home location, but principally research has investigated the way in which social connections between users can be exploited to understand location. The work of [11] provides the first serious attempt at this by using friendship connections on Facebook in order to locate users. They note that the likelihood of a friendship decreases as distance between two users grows. Therefore, if a user's location is unknown, the *known* locations of their friends can be used to infer the user's home location. Prediction accuracy increases with the number of friends. For users with 16 or more friends with known locations, 67.5% can be located within 25 miles, as compared to only 57.2% accuracy for an IP-based location. By removing 75% of known addresses, [11] presented a much more difficult evaluation problem. Although performance decreases, they are still able to achieve an accuracy of 57.4% within 25 miles. Work by [131] replicates [11] on Twitter data and report an average error of 426 miles. By incorporating the locations of user contacts they reduce that error to only 364 miles, although the top 60% of users are located within 6 miles. A similar stance is taken by [43] who make use of a small set of known user locations from GPS-tagged Twitter posts in order to estimate locations for users without any explicit location information. They take advantage of the social network and by jointly estimating the locations of a large number of Twitter users (≈100M) report an average error of 289km, better than previous methods [11, 131]. Unfortunately, their method is unable to make location estimates for users who do not have social ties to other known-location users. Social graph-based methods are not uniformly successful. Working with Sina Weibo data, [211] find they are able to locate user home locations with an average error of 789km, much larger than that reported in other work.

An alternative early paper focused on textual differences between regions by modeling the kinds of topics used on Twitter [68]. These latent, regional topics allow for only very rough geographic estimates across the US with an average error of 900km. By applying the task of user home location to Twitter data, [177] annotated geolocation information for 500 randomly selected users. Their approach does not rely on social information and instead aggregates the predictions from a number of variables including tweet message, location fields, time zones, and linked web

pages. Altogether their method is able to predict home location within a 100 mile radius for 79% of their users, which is quite impressive given the noise involved in Twitter data. The work of [33] learns a set of "local" vocabulary in an unsupervised fashion which they use to predict which of 5,913 U.S. cities a Twitter user lives in. This approach likewise ignores social information and is able to accurately classify 49.9% of users, an impressive feat for a purely lexical approach. Classifying user cities based purely on the use of specific learned keywords does poorly as reported by [103]. In their study they classify users from 10 distinct cities across the world but only achieve an accuracy of 65.7%.

More recently, [89] show that naive spatial clustering with k-means is able to locate users home locations within 50 kilometers for 86.2% of users. Importantly, only users with a certain number of geotagged posts could be classified. For other users, social clustering was still able to achieve 71.2% accuracy within 50 kilometers. While most home location detection papers have focused on locating users across a large area (e.g. the entire United States), [226] focus on Flickr users in New York City, classifying them into $100m^2$ grids. Because users post images on Flickr of many different types, they first identify images that are likely to have been taken at the user's home and then make predictions based on the geographic information in those images, classifying with 72% accuracy. Although all research on home location prediction suffers from similar pitfalls, including limited access to ground-truth locations, much progress has been made in recent years. While many users cannot be accurately predicted, even relatively simple clustering methods work for a majority of users. Because of limited ground-truth data, home location prediction has focused largely on a small number of users. Given the utility of social information in prediction, access to larger datasets might well lead to improved performance.

Predicting home locations is constrained by the number of users with known locations for evaluation. Predicting the location of an individual SM post is somewhat simpler in that there are many more geotagged posts. In the case of Twitter, even with only a small percentage of tweets containing location information, it is still possible for researchers to obtain millions of geotagged tweets for use in training and testing a model. Viewing the geolocation problem as one of predicting locations for individual SM posts also allows researchers to explore the ability of their models to forecast where an individuals' *next* SM post will be located. This can be done by taking advantage of spatio-temporal patterns of SM users individually or by comparing individual patterns with those of similar users, incorporating social dynamics.

As with home location prediction, [177] present initial tweet location results which are quite impressive. Their system aggregates location information from tweets, location fields, and other information such as websites and time zones mentioned in a tweet, achieving an average (mean) error distance of 1408km, but a median error of only 29.7km. This distinction between mean and median errors is driven by a small number of tweets which are very poorly localized, while the median error demonstrates that the majority of tweets are localized with high accuracy. While [177]'s positive results rely on the combination of a variety of data types mined from SM, [4] present a nonparametric Bayesian model which relies solely on the text of tweets. The model relies on the fact that vocabulary is shaped by geographic location and is able to achieve an average error of only 91.5km on predicting geotagged tweet locations. While different datasets make comparison difficult, this result certainly compares favorably against the 1408km average error of [177], showcasing the information hidden inside SM text. Unfortunately, learning with the model is not scalable to the massive quantities of available SM data. A much more efficient algorithm which is also based solely on text is presented by [208]. In contrast to [177] who rely on metadata specific to Twitter, this allows the model to be applied to a range of SM including Twitter, Wikipedia, and Flickr. They find promising results, with 49.2% of U.S. tweets located within 100 miles. For English Wikipedia data that number rises to 88.9% and for Flickr images, which are located only based on user-supplied image tags, that number is 66.0% [208].

These studies have all focused specifically on nowcasting the location of SM posts, and largely ignore historical patterns of individual users. This renders their techniques unsuitable for forecasting, which requires an understanding of how user locations change over time. This type of

SM mobility modeling has received much recent attention as well as impressive results. A move in this direction is made by [40] who examine Sina Weibo, a Chinese equivalent to Twitter. They use tweet texts to understand individual user interests and then use these interests to predict locations. Although their model could be used to forecast, they evaluate it only on nowcasting tweet locations as in the previous studies. Direct comparison cannot be made with other studies given that they restrict their data to a single city, Beijing. They classify 72% of tweets within 1km, but their method excludes any tweets for which no classification was made, giving their model a recall of only 15.8%. A more useful model is presented by [201], which incorporates knowledge about patterns of human movement, namely the fact that individuals rarely change their patterns and that individuals with similar social backgrounds tend to share patterns. They use bus and taxi GPS data in order to model general spatial patterns of individuals in Beijing and incorporate geotagged SM tweets to understand a user's movement patterns. Their model produces a ranked list of possible locations and they present their results as Acc@topP, representing the accuracy of predictions so long as the actual location is within the top P predictions. Roughly 50% of tweets are predicting within the top 60 locations, out of a total possible of 118,534.

Recent work has expanded on the possibility of incorporating knowledge of social dynamics to predict user movements. The work of [219] specifically treats each user as a member of a latent group, where each group shares a set of movement patterns. A soft assignment for each user is made, allowing a user to belong to multiple groups simultaneously. Again, they model users on a city level both in NYC and Los Angeles. Modeling multiple groups improves accuracy by 12.7% over a model which treats all users as members of a single group.

There are a number of hurdles to overcome in geolocation prediction. Perhaps the largest is the small percentage of geotagged data for use as a ground-truth. This is especially problematic for models that predict individual user movements, since users that do post geotagged information may well differ systematically from other SM users. A further difficulty in comparison is the variety of evaluation metrics presented, as well as differences in scale. Some studies make predictions at a country- or world-level, while others predict movements within an individual city. Errors acceptable at one level may be unacceptable at another and differences between corpora may make prediction arbitrarily easier or more difficult based on properties that are not well understood. Current efforts clearly indicate that geolocation prediction is possible. Incorporating social demographics in order to pattern users together appears to greatly improve performance [40, 201, 219]. When available, metadata are powerful sources of geolocation information [177], but text alone can achieve impressive results due to differences in vocabulary across geographic regions [4]. If future work can leverage these initial findings and work with larger sources of data then accuracy for geolocation will likely improve greatly.

Demographics

While geographic location is one important hidden property of SM users, there are any number of demographic or psychological properties which can be inferred either from a user's SM relationships or posts. The two most studied hidden properties are age [51, 137, 154, 173, 194, 195] and gender [16, 25, 51, 72, 173, 194, 195, 215]. Techniques for discovering these properties have improved greatly over time and reveal the power of text analysis as well as social relationships in revealing hidden properties. Further work has explored a number of other latent properties such as race [16, 51, 137], education level [195], political affiliation [137, 197], occupation [92], income level [159, 195], and even willingness to volunteer [182]. While none of these tasks can be thought of as forecasting, the ability to infer the real-world properties of SM users is an important cornerstone towards expanding forecasting models.

A wide variety of studies have investigated the ability of SM data from various platforms to infer user gender, all with positive results. Gender detection is typically formulated as a classification problem with two outcomes, "male" and "female". Initial work by [25] was somewhat pessimistic. Trained on text from tweets they achieved an accuracy of only 74.5%, better than baseline (54.3%) or human (68.7%) performance. Incorporating a user's full name,

along with other metadata, increases performance to 91.8%, while [16] report 90.2% accuracy on their Twitter dataset. The work of [72] examines Youtube comments as a source of gender information and achieve high accuracy (89%), although gender imbalances in their dataset lead to a much higher baseline performance (70%). Importantly, they note that inference is much easier for individuals who tend to associate within their own gender (94%) than for users who associate with the opposite gender (47%). This implies that gender cues in language are used differently based on social factors and may inherently limit the accuracy achieved from text alone. In spite of this issue, text-based techniques have made large strides. The work of [173] presents a weighted lexicon-based approach that achieves 91.9% accuracy on Facebook and 90.0% accuracy on Twitter data. The work of [194] reports their results in terms of receiver operating characteristic (ROC) analysis, scoring 0.90 for text alone and 0.95 when sentiment analysis is also incorporated. Even without using text directly, gender can be inferred to some degree. The work of [215] uses image-based topics from Pinterest to achieve 71.9% accuracy, while [195] achieve an AUC of 0.76 using inferred user interests. The work of [51] infers gender by fitting users to match the demographics of websites they follow on Twitter and similarly achieve an F1-score of 0.75, indicating that gender can be inferred even when text is not available.

Age detection is also a useful tool for SM prediction tasks. While research is consistent in attempting to infer only two gender categories, work on age detection is split between inferring age in terms of integer values (e.g., years) or as a regression task versus breaking age into brackets for multi-category classification. For instance, [152, 194, 195] attempt to infer whether a user is above or below 25 years in age. The high F1-score of 0.88 reported by [152] is limited by the fact that they train a model using only word n-grams based on Dutch SM posts. While accurate, such methods are known to be brittle and do not generalize well to out-of-domain data such as posts from other SM sites. An AUC of only 0.66 is reported by [195], much lower than the score of 0.90 they achieve for gender detection. Age detection can be improved by adding more detailed sentiment analysis as in [194], who improve their AUC score to 0.83, still much lower than the score of 0.95 for gender. This difficulty is reflected as well in [137] whose model based on general demographics performs very poorly on the above/below 25 classification task. They find best performance for a simple logistic regression which achieves an accuracy of 83.3%. The work of [51] also attempts to use outside demographic information to improve performance and find somewhat more promising results. They split age into brackets of 10 years (e.g., 25-34, 35-44) and report correlation coefficients between inferred and actual ages in each bracket. They find best performance for ages 18-24 ($r = 0.78$) and worst for ages 35-44 ($r = 0.55$). Both [173] and [154] report mean errors in terms of years. The work of [173] reports a correlation coefficient of 0.83, better than [51], with a mean error of 3.42 years for Facebook users and 3.76 years for bloggers. Network analysis between friends on the Slovenian SM platform POKEC is used by [154] to infer age. Their method has lower accuracy ($r = 0.70$, MAE = 4.15 yrs) but their results rely on only 5% of social network users having a known age. Instagram tags and profile features are investigated by [88] who report good accuracy 79.5% when making use of both sets of features. Additionally, they evaluate their models on a completely held out set of users, ensuring the generalizability of their method.

Beyond gender and age there are a number of other demographic variables which might be useful features for forecasting. The work of [195] refers to these properties as *psycho-demographics*. While there are any number of attributes which could be analyzed, we focus our review here on six properties: ethnicity, income, education level, occupation, political party affiliation, and willingness to volunteer. These six attributes give an idea of what can be accurately inferred from SM data both in terms of traditional demographics, but also in terms of behavioral patterns. Due to the sparsity of work on any particular topic, we make no general claim as to the feasibility of making specific psycho-demographic inferences and instead argue that psycho-demographics appear to be detectable generally with a moderate level of accuracy.

For example, consider the case of detecting ethnicity. Early work demonstrated a promising accuracy of 81.3% across 4 ethnic categories, but relied on having a user's first and last names

which cannot always be assumed from SM data [16]. Later work showed that a much lower level of accuracy ($\sim 60\%$) is possible without any kind of text data or names, based solely on matching user demographics against the demographics of websites they follow on Twitter [51]. The work of [137] incorporates general demographic trends including a user's first name and reports accuracy of 82.3%. Research on inferring user income has seen similar results. By examining a number of features to infer user income, including other psycho-demographic traits, [159] achieves a correlation of $r = 0.633$ with a mean absolute error (MAE) of £9535. The work of [195] presents their results only in terms of AUC, making comparison difficult, but they achieve an AUC of 0.67 using only inferred user interests, and an AUC of 0.73 using more traditional text analysis. In the same study, [195] also model education levels with slightly better results, AUC = 0.70 for user interests and 0.77 for text analysis. Political party preference can also be treated as a latent psycho-demographic attribute, classifying users as either likely Republican or Democrat voters. The work of [197] reports accuracy of 81% when using user text and neighbors. Occupation prediction shows similar moderate inference success, with an F1-score of .78 across 8 broad occupation labels [92]. Willingness to volunteer proves to be somewhat easier to predict using data from the SM platform LinkedIn, with an F1-score of 0.87 [182], and increases to 0.899 when fusing multiple SM source [96].

Inference for user pyscho-demographics shows a wide range of levels of success. Social media *can* be used to infer user attributes but the level of success depends on the type of data used and the difficulty of the inference task. The highest level of performance is seen in predicting gender, where text analysis of SM posts has reached similar levels of achievement to models based on user first and last names [194]. At the same time, text alone is limited based on the fact that some users systematically pattern with the opposite gender [72]. Detecting other user traits shows much more moderate levels of performance, but the sheer variety of traits which have been successfully inferred demonstrates the amount of information hidden in SM behavior. Because psycho-demographics can be used to improve performance on actual prediction tasks, even moderate success in this area is promising.

Table 5. Summary of Studies on User Characteristics

Article	Topic	Data Source	Data Size	Features	Task	Success Rate
Perozzi [154]	Age	O	1.6M users	Social	Regression	$r^2 = 0.49$
Zhang [220]	Age	T	55K users	N-gram, Social	Classification	F1 0.81
Han [88]	Age	I	20K users	Semantic, Social, Behavioral	Classification	Acc. 78%
Peersman [152]	Age	O	1.5M posts	N-gram	Classification	F1 0.88
Ardehaly [137]	Demographics	T	18M tweets, 2.7M users	Metadata, N-gram, Social	Classification	Acc. 75% (age), 82% (race)
Bergsma [16]	Demographics	T	168M users	N-gram, Social	Classification	Acc. 90% (gender), 85% (race)
Culotta [51]	Demographics	T, O	46K users	Social, Non-SM	Classification	F1 0.75 (gender), 0.69 (ethnicity)
Volkova [194]	Demographics	T	24.9M tweets	Semantic, Social	Classification	AUC 0.83 (age), 0.95 (gender)
Volkova [195]	Demographics	T	4K users	N-gram, Social	Classification	AUC 0.66 (age), 0.90 (gender)
Burger [25]	Gender	T	213M tweets, 18.5M users	Metadata, N-gram	Classification	Acc. 92%
Filippova [72]	Gender	Y	6.9M users	N-gram, Social	Classification	Acc. 90%
Sap [173]	Gender	T, F, O	75K users	N-gram	Classification, Regression	Acc. 92% (gender) $r = 0.83$ (age)
You [215]	Gender	T, P	243K images	Semantic	Classification	Acc. 72%
Li [118]	Gender	SW	25K users	N-gram, Semantic	Classification	Acc. 94.3%
Eisenstein [68]	Geolocation	T	9.5K users	Semantic	Regression	MAE 900km
Ahmed [4]	Geolocation	T	573K tweets	Semantic, Spatial	Clustering	MAE 91.5km
Backstrom [11]	Geolocation	F	2.9M users, 30.6M edges	Social, Spatial	Regression	Acc. 57%
Chang [33]	Geolocation	T	4.1M tweets	N-gram, Spatial	Classification	MAE 509mi
Chen [40]	Geolocation	SW	1.1M posts	Semantic, Spatial	Classification	Acc. 70%
Harrison [89]	Geolocation	T	580K tweets, 245K users	Spatial	Clustering	Acc. 86%
Schulz [177]	Geolocation	T	80M tweets	Metadata, Semantic, Non-SM	Regression	MAE 1408km
McGee [131]	Geolocation	T	73M users	Social, Spatial	Regression	MAE 364mi
Compton [43]	Geolocation	T	111M users	Social, Spatial	Regression	MAE 289km
Xu [211]	Geolocation	TW	2M users	Social, Spatial	Regression	MAE 783km
Kinsella [103]	Geolocation	T	Not specified	Semantic	Classification	Acc. 65.7% (baseline 40.3%)
Wang [201]	Geolocation	SW, O	7.3M check-ins	Spatial, Non-SM	Classification	Acc. 45%
Wing [208]	Geolocation	T, W, FR	38M tweets, 864K posts	N-gram, Spatial	Classification	Acc. 49% (T), 89% (W)
Zheng [226]	Geolocation	FR	48K images, 192 users	Spatial, Visual	Classification	Acc. 72%
Zhang [219]	Geolocation	T	1.3M tweets	Semantic, Spatial	Classification	Acc. 50% (LA), 38% (NYC)
Preoctiu [159]	Income	T	10.8M tweets	Metadata, N-gram, Semantic	Regression	$r = 0.63$
Hu [92]	Occupation	T, L	9.8K users	N-gram, Semantic	Classification	F1 0.78
Volkova [197]	Pol. party	T	300 users	N-gram, Social	Classification	Acc. 81%
Song [182]	Volunteerism	T, F, L	5.9K users	Metadata, Semantic, Social	Classification	F1 0.87

T = Twitter, F = Facebook, W = Wikipedia, SW = Sina Weibo, TW = Tencent Weibo, L = LinkedIn, P = Pinterest, I = Instagram, FR = Flickr, Y = Youtube, O = Blogs, other

Discussion

Can Social Media be Used to Predict the Future?

Having thoroughly reviewed the literature in previous sections, we find strong evidence to support the notion that SM can be used not just to detect current real-world events, but also to make accurate forecasts into the future. Great strides have been made in the literature since the last set of reviews on SM prediction [100, 176, 216], and in that time the cautious optimism of previous reviews has largely been borne out. Positive results have been found in every area examined, although the degree of success is heavily moderated in part by a number of factors which we will later address (Q2). While the reviewed literature likely suffers from publication bias, failing to include many studies which found no predictive power but were not published, the fact that so many positive results have been published indicates that in some cases SM does carry great predictive power.

While there is no doubt that progress has been made, SM forecasting largely faces the same set of challenges which had previously been identified. First, any predictive signal in SM is surrounded by large quantities of noise [13, 67] and extracting meaningful signal is no trivial task. Further, SM data is biased [155, 167]. Although we can think of SM users as sensors of the real world [47, 62, 168], it is well known that SM users are not representative of the total population [155] and although many SM posts describe what is going on in a person's daily life [95], these posts are not necessarily representative of everything going on in the world around the user. As noted by [100, 176, 204, 216], research in SM is marred by issues of generalizability, research which makes accurate predictions on one data set may not prove useful in another. This issue is then compounded by the use of powerful data-driven modeling techniques without the use of any domain knowledge to link predictive power to underlying mechanisms. This is especially important given that many of the tasks researchers hope to predict are fundamentally complex phenomena.

The good news is that researchers in many different areas have identified at least partial solutions to many of these problems. In spite of these difficulties, they have achieved moderate accuracy of forecasts across a wide range of predictive tasks. The bad news is that achieving positive results is not necessarily straightforward. Addressing these major problems may require explicitly modeling for user biases, applying complex data-driven models, training on varied data sources, and incorporating domain-specific knowledge and theory into the modeling process. In the following section, we tackle each of these difficulties in turn, identifying studies that illustrate best practices in SM prediction.

What leads to Social Media Prediction Success?

Given the variety of results presented, it is necessary to identify general trends which differentiate SM prediction success from failure. Previous literature reviews have had very little to say on the topic. The work of [100] notes that studies which use advanced techniques for filtering SM data based on keywords are more likely to find predictive success than those who use simpler filters. Further, they find that sentiment analysis techniques were highly controversial, in some cases lending no predictive power and in others proving highly useful. Based on an alternate review of the literature, [176] predicted that data-driven statistical models would increasingly find a place in SM forecasting, which has been the case.

With the explosion in research on SM forecasting over the past few years, we can make somewhat more specific statements regarding what practices researchers have used in order to find positive results. We identify four major issues which SM researchers must confront and make suggestions regarding best practices for each. We discuss each issue and best practice in detail below, presenting concrete examples demonstrating their importance for future work. The identified best practices include:

1. Applying appropriate techniques to overcome noisy data

2. Explicitly accounting for SM data biases
3. Learning from heterogeneous data sources
4. Incorporating domain-specific knowledge and theory

Overcoming Noisy Data

While one advantage of SM data is the large quantity of data generated by users, researchers should also keep in mind that not all of that data will be useful for any particular forecasting task. Consider that in their study of adverse drug reactions [18] find a total of 239 potential cancer drug users narrowed down from a total set of 2 *billion* tweets. While this may be an extreme example, the large degree of noise—specifically information unrelated to the prediction task—represents a significant obstacle which SM researchers must overcome.

Data filtering as an important step the data analysis process was identified by [100]. Of the studies they review, papers which manually selected keywords for filtering data supported SM predictive power only 50% of the time. In contrast, every paper that used statistical algorithms to select keywords automatically found positive SM forecasting results. The work of [190] provides a possible explanation for this, noting that pre-selected keywords such as hashtags, which might at first glance appear reasonable, can easily lead to poor real-world predictions. For instance, forecasting protest volumes based on hashtag usage would have led to very poor predictions in the case of Turkey's 2013 Gezi Park protests. While hashtag use dropped sharply, protest volumes and talk on SM related to the protests actually increased, because the protests were so large users no longer felt the need to use hashtags to coordinate [190]. This highlights the danger of focusing heavily on individual hashtags or keywords within a SM landscape which changes rapidly [204]. Without a principled method for filtering data, researchers risk creating models that fail due to minor shifts in word usage.

A further challenge for filtering comes from studies that make now- and forecasts based on a set of users who have been filtered for specific qualities. This type of filtering can bias predictive systems based on what was filtered [190]. For instance, consider the work of [196, 197] who predict political party preferences for users who are either self-labeled as Democrats or Republicans or who follow exclusively Democrat or Republican candidates on Twitter. Filtering in this fashion ensures that ground-truth labels are more accurate, but also creates a bias since most Twitter users neither self-label with a political party or follow political candidates [42]. Another example of poor user filtering comes from [220] who infer a user's age from Tweets wishing them a happy birthday while mentioning their age. Because older users rarely have their exact age mentioned, their model performs well only for users aged $14 - 22$.

An alternative approach is to use statistical techniques in order to identify signal within the massive amount of noise generated by SM without overtly removing data as in the case of filtering. Models capable of dealing with such large amounts of data can take advantage of non-obvious relations. For instance, [109] use weather-related keywords in order to predict rainfall. This allows them to learn relations between users tweeting about words such as "rain", "sun", and even indirect words such as "beach" or "tv", which might indicate activities individuals are engaging in during a sunny or rainy day. Because they rely on a fixed list of keywords, the scope of these indirect signals is quite limited. While many users tweeting the word "fireplace" might be a useful cue to colder weather, because it is not included as a keyword it is effectively ignored. A much more rigorous approach is examined in [122] who nowcast user stress levels based on a corpus of one billion tweets. Using a neural network-based architecture, they include both a set of features using stress-based keywords, but also learn from every tweet based on meanings learned from all of its words. They find that in isolation the keyword-based features are more effective than learned tweet meanings, but the combination of both significantly improves nowcasting results. In a similar fashion, [194] build upon their previous work in demographic inference by removing the need for user or keyword filtering. This allows them to learn relations between user demographics and Twitter data without having a set of biased users as in [196, 197].

Noise may also be reduced in SM text data through the use of various natural language processing (NLP) algorithms. In particular [13] examine types of noise in SM data and the effect that various NLP algorithms have on reducing it. They conclude that while traditional NLP methods developed for non-SM data are not ideally suited for SM text, they should still be generally effective. In practice, these signal extraction methods, which do not rely on keywords, are not always perfect. Consider the case of [222] who use a number of NLP techniques to reduce linguistic noise and then train a classifier to identify asthma-relevant tweets. This extraction technique does not rely on keywords and significantly improves correlations between tweets and asthma prevalence as well as monthly hospital visits. On the other hand, the method actually *reduces* correlation between tweets and daily hospital visits. Even with advanced statistical models, they have difficulty forecasting whether they will see a "high" or "low" level of hospital visits either in the next day or week, reporting only 66% accuracy for the following day. As [67] describe in great detail, standard methods for extracting signal from text are often poorly suited to online SM data. Researchers focused on SM data should therefore be wary of blindly applying such methods which were originally developed for text written by traditional media or scholarly sources.

While SM data is inherently noisy, researchers have found a number of techniques to reduce this noise and extract the signal necessary for various now- and forecasting tasks. In some cases, simple keyword or hashtag filtering is sufficient, but researchers should be aware filtering is best done in a principled fashion [100]. Keywords chosen based on researcher intuition may be fragile, both removing possibly important information and focusing on signals which may change in importance as particular words and hashtags rise or fall in popularity [190, 204]. Alternatively, statistical techniques for signal detection may be utilized without filtering. These methods automatically infer which signals are important, bypassing the need for potentially biased researcher judgments, but they have to confront the issue of noise, which may make learning difficult. In many cases, statistical methods have shown great promise, but powerful statistical models do not, in and of themselves, guarantee predictive power.

Accounting for Data Bias

As noted in our discussion of study bias, the use of SM data to predict real-world attributes introduces a host of biases with which researchers must contend [64, 155, 167, 176, 190]. Not only do SM users differ from the general population [64, 155], but the content of SM posts also may not reflect every aspect of the real world [167]. These issues have been widely discussed within the literature and we broadly consider them under the name of *SM data bias*.

A recent Pew Research Center study [155] shows that, apart from age, general SM adoption is largely commensurate with the growth of general Internet use among U.S. adults and representative of the U.S. population by both gender and ethnicity. In a separate Pew Research Center study [63] however, we find that average SM usage is not as consistent across demographics on a *per-platform basis*. This is especially relevant when SM research is constrained to a single platform (which we often find in past studies, as shown in the tables above). As of 2015, among all Internet users, women are better represented than men on Facebook, with 77% of Internet-using women being on Facebook versus 66% of Internet-using men. Pinterest usage is considerably more shifted toward female than male users (44% versus 16%). Conversely, men are more likely to be found making use of discussion platforms such as Reddit or Digg, at 20% of male versus 11% of female Internet users. Platforms like LinkedIn and Twitter see much closer utilization rates between men and women, but exhibit much sharper socio-economic disparities [63]. 41% of Internet users with an annual income over $75,000 use LinkedIn, while only 21% of the same income bracket use Twitter. In terms of race, 47% of African-American Internet users utilize Instagram compared to 21% of white Internet users; this dichotomy is switched on Pinterest, where 32% of white and 23% of African-American Internet users maintain accounts.

These population biases like age, gender, ethnicity, and socio-economic status are potentially critical to researchers in election prediction and public health. Indeed, [75], appealing to the

"Literary Digest" poll of 1936, highlights the dangers of heavily skewed sample populations when attempting to make statements about election outcomes when the demographics of those who use SM may differ significantly from the demographics of those who are likely to vote (e.g., seniors, higher educational attainment). Some election studies have taken note of this: [191] observed that although sampled demographics were heavily skewed, the authors make no attempt to correct this bias and instead suggest the sample may be a good representative population of "opinion leaders". The work of [172] attempts to rectify known population skew using a weighting scheme to more accurately reflect the true electorate population. Many studies [74, 179, 181], however, make no mention of this potential bias.

Sampling strategies themselves can also be problematic. For research involving voluntary data-sharing (e.g., [53, 54]), recent work has shown that self-reported Internet use is generally unreliable [175]. Additionally, [170] show that not only do SM users' network of real-world interactions differ from their SM interactions, this difference is typically larger than that perceived by the user. Keyword search methods of data sampling are subject to numerous linguistic factors that lead to bias [190, 192]. The work of [182] for example, attempts to predict a LinkedIn user's willingness to volunteer with some apparent success, but their model is trained on users who include "volunteer" in their profile in addition to all of those user's LinkedIn connections. The model is biased toward identifying not just volunteers, but volunteers who choose to self-identify themselves as such. When attempting to construct friend-follower graphs, connectivity-based sampling methods lead to unrepresentative sample populations while random population sampling leads to erroneous connectivity traits [116].

And biases inherent to the data analyzed are not the only danger for SM researchers. Creating a predictive model entails learning a relationship between the training data (from SM) and some future real-world attributes. The advantage of SM data is often its large quantity, which lends itself well to statistical modeling. At the same time, inappropriate use of statistical techniques can easily lead to a problem known as *overfitting* [91]. Overfitting occurs when a statistical model does not just learn the target relationship, but also captures the peculiarities and randomness inherent to the data. An overfit model may perform well under k-fold cross validation, but real accuracy will suffer when confronted with true hold-out validation data. For example, Google FluTrends showed incredible promise in the early stages of research [45]. However, the model exhibited classic symptoms of overfitting, where model performance suffered greatly once confronted with new data [148].

Because an overfit model cannot necessarily be detected based solely on its results, researchers must preemptively take measures to ensure their models will be able to generalize to novel data. For instance, in predicting depression in SM users [53, 55] apply principle component analysis (PCA) to determine relevant data features, which helps in preventing overfitting by eliminating feature redundancy in the data. Overfitting can also be combated through the use of certain model training schemes, as in [110] (via special choice of regularizers for outlier data) and discussed by [174, 183] (via dimension reduction).

These concerns, however, are often only paid lip service and precautions are taken infrequently [74, 167]. For example, in [191], the authors indicate that 4% of users were responsible for 40% of the sampled data but no caution was expressed for the possibility that the model might overfit, learning predictions based on a small handful of users who may not generalize well to other elections. In the case of Twitter, given that many studies search for specific keywords or users with specific characteristics, features may be very highly correlated. In these cases, training data itself can invite bias if the models that utilize them are reinforced by too much repetition. Because SM modeling is almost entirely data-driven, overfitting should be a constant concern.

Although general SM data biases are a consistent issue in the field, there are a number of possible methods researchers can use to ensure the robustness of their results. Demographic biases should be quantified and accounted for whenever possible. Filtering based on keywords and user characteristics are easy techniques to reduce data noise but introduce biases and should be used sparingly, replaced by alternative noise reduction methods discussed in the previous section.

Finally, model overfitting should be recognized as a serious possibility and avoided through the use of certain model training schemes. Accounting for these biases not only leads to a better understanding of underlying processes in SM, but also helps ensure models can be applied beyond the data they were trained on.

Improving Generalizability

One of the biggest concerns for SM data biases is that results which are successful in one context may fail in others. A good forecasting model should make accurate predictions not just on data from today, but also on data from tomorrow. If a forecasting model can only be applied to data from one particular location in one particular year, the model lacks the ability to forecast in a useful fashion. To accomplish this, the model needs to have learned a relationship between SM and real-world events which are unlikely to change over time. In essence, a model needs to be robust enough to generalize to novel data, not knowing ahead of time what that new data might look like. Failure to generate robust predictions might lead to poor performance whenever major changes in the world occur, often exactly when accurate predictions are most valuable. The ability to generalize findings should thus be of primary interest to researchers attempting to forecast the future, and while data biases and overfitting are two factors which can lead to generalization issues they are far from the only concerns.

The most basic cause for model generalization problems comes from a mismatch between the data a model is trained with and the data which will eventually be used in forecasting. As noted previously, it is not always possible to know what this mismatch might entail, but there are a variety of common issues that researchers can attempt to address. In particular, we note cases where poor performance is obtained due to usage of data which is too narrow in scope, where models are not evaluated across multiple possible domains of interest, and where learning occurs with data that comes from a narrow window in time.

Learning from a very narrow set of data is a large problem within the field of SM research. By narrow, we mean simply that the data may be insufficiently diverse to allow a model to perform well at a wide range of conditions. For instance, consider the case of nowcasting user political preferences on Twitter. If the goal is to detect preferences for all users, then the methodology of [197] would be inappropriate. In their study, they trained using a small set of users who had self-identified as either Democrats or Republicans, which may not be representative of general users on Twitter. Similarly, the methods of [56] in forecasting protest volume rely on a set of hashtags which were identified after the fact by mining Wikipedia pages. Applying this work to future events would require additional methods for automatically identifying the sets of hashtags relevant to an event, but [190] warn that any analysis based on filtering for hashtags or keywords is likely to produce a dataset with particularities specific to those keywords which may not be generalizable to future events.

The possibility of very fragile models based on keywords alone is demonstrated in [191]. They report that the simple percentage of mentions on Twitter for political parties in Germany reflects the share of the vote each party will win. While a very exciting premise, as [99] point out, the results are not generalizable and rely critically on excluding the German "Pirate Party", which was mentioned on Twitter that year more than any other political party yet garnered only 2.1% of the vote. Changing the days of data collection likewise had major effects on the election forecast, indicating poor robustness.

One way for researchers to overcome these issues of narrow data involves learning from data over multiple SM platforms. This strategy has been used very successfully in the area of user demographic nowcasting, where relationships between user demographics and SM behavior might vary from platform to platform. For instance, [173] infer a user's age and gender based on the words they use, learning from Twitter, Facebook, or blogs either separately or together. A model trained on Facebook alone will perform well on data from Facebook, but does much more poorly on data from other sources. Models trained on a variety of sources may perform more poorly on data from any individual SM platform, but the results are more robust to changes in data source.

In the work of [96, 182] an argument is made that integrating information from multiple SM platforms increases robustness of results and further has the benefit that information from different platforms is often complementary. User profiles on LinkedIn, for instance, generally contain information on educational achievement while Facebook users are much more likely to list their gender [182].

Text-based geolocation from data on a variety of SM platforms and languages is likewise explored by [208]. Because their corpora are not equatable, they cannot learn over all the data at once, but their results demonstrate the utility of including more than one evaluation. They posit two models, which perform roughly equally well on data from Twitter. If they had only examined Twitter data, as in most studies on SM forecasting, they might reasonably have chosen the simpler of the two models. Unfortunately, this simpler variant performs much worse on data from Wikipedia and Flickr, a finding which would not have been recognized without investigating multiple data sources. A similar case of evaluating on multiple data sources can be found in detecting and forecasting depression [52, 54, 55, 57, 188]. While these studies differ slightly in their methods, they make use of the same basic text features [153]. The methodology has found positive results using Twitter [52, 55, 188], Facebook [54], and Reddit [57] and while most work has been on English-speaking SM users, [188] validate the findings for Japanese-speaking users as well. The method has found success in detecting depression [55], but also for forecasting postpartum depression [52, 54] and forecasting whether a depressed individual will start thinking about suicide [57]. By validating the methodology on such a wide variety of data sources, researchers can feel more confident that the same methods will be useful when applied to similar prediction tasks.

Incorporating non-SM data into forecasting models can also be a useful tool for increasing model robustness. By linking the model to data which is often less noisy and whose relation to prediction is better understood, studies which incorporate non-SM data are often able to achieve better forecasting performance. This practice is more common in some areas than in others. For instance, to detect adverse drug effects researchers make use of extensive domain-specific knowledge about drug side effects from non-SM databases [18, 71, 178, 214]. Incorporating non-SM data in this case makes side effect identification much simpler, but does not preclude models from detecting side effects which were previously unknown [71, 214], an important step for public health researchers. Similarly, non-SM data can be used to enhance demographic prediction by matching unknown users against known demographic patterns [50, 51], geolocation by modeling population-level traffic patterns [201], or by mapping place names to locations [177] and election prediction by factoring in variables known to affect election outcomes [206].

Lastly, in many cases it is important that forecasting models be trained on data that is representative of the timescale being used for prediction. For instance, [109] use tweets from five cities in the UK to predict rainfall. Because they only have one year's worth of data, the model is always trained on 10 months and asked to predict rainfall for the remaining two. Even with such limited training data, the model does quite well, with the notable exception of predicting the weather in July, which is both a summer month, with individuals tweeting about sunny, outdoor activities, but was also the second most rainy month in the dataset. If training had utilized a longer timespan of data, such regular annual patterns could be forecast more easily. The same can be said of election outcome predictions, where training typically occurs for a single set of elections [27, 38, 164, 172, 179, 181, 191], which may or may not be representative of other elections past and future. Studies making use of data from multiple elections will be necessary in order to make statements regarding election forecasting that might be more reliable.

In summary, researchers should constantly be aware that decisions they make regarding training data and evaluation are likely to have a considerable impact on their ability to forecast for particular use cases. We advise researchers to make use of their domain expertise in order to determine what aspects of their model are most in need of generalization. Training using data from multiple SM platforms may be wise for a task such as protest forecasting where protesters may in future adopt a platform other than Twitter, but may be unwise for a task where only a single platform is required and is unlikely to be replaced. Training using data over an extended

period of time may be appropriate when temporal patterns are relevant as with rainfall, but may be inappropriate if historical patterns have changed so dramatically as to be irrelevant. Although there is no one solution to problems of model generalization, future researchers would do well to consider how the general guidelines presented here might apply for their own use cases.

Incorporating Domain-Specific Knowledge

Research into SM forecasting has largely found success thanks to robust statistical models which take advantage of large quantities of SM data [176]. As we mentioned previously, applying canonical machine learning models can help researchers overcome the tremendous noise in SM data, but leaves open the possibility of overfitting, especially when data of only one type is used. Careful choice of data sources, model techniques to account for biases, and evaluation on multiple data sets can all help to overcome some of the limitations to these types of models. An additional avenue which has seen great success in many areas of SM prediction is the use of domain-specific knowledge in order to augment statistical models.

By domain-specific knowledge, we mean here the knowledge and theory specific to a particular task or field which has been validated by existing research. By incorporating patterns that are already known, researchers can point their statistical models in the right direction. Not only can this improve model results, but it can also help to ensure generalizability. For instance, consider the detection and forecasting of depressive symptoms in SM users. A great deal is known from psychology about various types of depression and this knowledge has largely been incorporated into existing predictive research [52, 54, 55, 57, 188]. In addition, the choice of ground-truth data—study volunteers and random control data in the case of [52, 54, 55]—and how to manage the task of validation—confirmation bias and self-reporting—is well understood. Although all of this work relies on relatively naive text analysis [153], research has been generalized well across datasets. Knowledge from psychology identifies the underlying behaviors which are linked to depression and which are expressed in SM usage. Because researchers are able to take advantage of this knowledge, they can achieve reasonable accuracy even with naive methods that do not take advantage of the scale of SM data.

In the case of depression, domain-specific knowledge manifests itself in terms of choosing model features which are linked to the target prediction of interest. Domain knowledge can also be incorporated into the structure of the forecasting model itself. Much work in geolocation now- and forecasting is built upon the knowledge that individuals tend to revisit locations they have been to before [201]. This fails, however, to take advantage of the patterns between individuals which exist. Results for SM users with little historical data can be improved by assuming they are similar to the general population [221]. The best results in the field come from models which specifically attempt to model what is theoretically known as *homophily*, individuals who associate with one another are more likely to share travel patterns [11, 177, 219]. Finding ways of incorporating sociological knowledge has allowed researchers to greatly improve location forecasts in spite of the fact that most location information in SM is quite sparse [40] and the fact that physical, daily interaction networks and virtual SM interaction networks differ greatly from one another [65, 170].

Work in demographics prediction from [1, 196] likewise use a theory-driven approach, using social influence theory in order to construct scalable network features for accurately detecting user preferences on Twitter. Indeed, the authors note that many classical machine learning models, such as logistic regression, fail to capture interrelations between users, and instead represent each user or tweet as an independent instance within the data. However, this clearly glosses over the network structure between Twitter users following and mentioning one another. The authors leverage this domain knowledge by assuming that users that follow one another likely have similar interests, and more specifically that influential users can be used to identify sub-groups of users that likely have similar interests through following or other interactions.

Failing to incorporate domain-specific knowledge or capture known or even hypothesized dynamics within the physical system of interest has been blamed for a growing number of bad

outcomes in SM prediction research. Likely the most noteworthy example is Google Flu Trends [111]. Although using relative search volume for symptoms of influenza and influenza-like-illness seems logical, there is little to no theoretical basis for this relationship. Indeed, following the same logic in terms of search volume, stock market prediction faced the same problem when [14] showed that the words "colour" and "restaurant" were the second and third best search term predictors of stock market movements. The authors had no theoretical basis on which to conclude that these results could be valid beyond what naive data analysis had shown them. Even the best predictor, "debt", was not entirely clear to the authors as to why it performed so well.

Conclusion

In this systematic literature review we examine the ability of SM data to forecast real world events and characteristics across a variety of disciplines. We have focused our review toward answering two questions: Can SM be used to predict the future, and if so, how is this best accomplished?

First, the good news: in addressing our first research question, we find that SM data has been used to make accurate forecasts across all of the disciplines examined. Additionally, topics that can be shown to be directly relevant to SM users and how they interact with SM make more successful predictions, such as user location, user demographics, and civil unrest. The bad news: in addressing our second research question, we detail four major pitfalls which have made SM prediction difficult. Noisy data, SM data biases, lack of generalizability, and difficulty incorporating domain-specific knowledge and theory lead to a fundamentally complex prediction task.

For each of these pitfalls, we examined the literature to find papers which best overcame these difficulties identifying best practices. These include, but are not limited to 1) carefully filtering out irrelevant information, such as by learning appropriate keywords [13], 2) incorporating known SM data biases by, for example, factoring in the effect of skewed demographics, 3) avoiding overfitting models to ensure predictions will be robust to future data by only incorporating relevant data features during model training such as in [54], as in [172], and 4) appealing to domain knowledge and theory, potentially through validation studies like [170]. By following these best practices, future researchers will better be able to make use of SM data, avoiding mistakes in past research which have led to poor performance.

References

1. M. A. Abbasi, J. Tang, and H. Liu. Scalable learning of users' preferences using networked data. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 4–12, New York, NY, USA, 2014. ACM.

2. H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013.

3. F. Abel, E. Diaz-Aviles, N. Henze, D. Krause, and P. Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pages 276–280. IEEE, 2010.

4. A. Ahmed, L. Hong, and A. J. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36. ACM, 2013.

5. M. Alanyali, T. Preis, and H. S. Moat. Tracking protests using geotagged flickr photographs. *PloS one*, 11(3):e0150466, 2016.

6. N. Alsaedi, P. Burnap, and O. Rana. Identifying disruptive events from social media to enhance situational awareness. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 934–941. IEEE, 2015.

7. N. Alsaedi, P. Burnap, and O. Rana. Sensing real-world events using arabic twitter posts. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

8. E. W. Anderson. Customer satisfaction and word of mouth. *Journal of service research*, 1(1):5–17, 1998.

9. S. Asur and B. A. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

10. M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi. Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1749–1758. ACM, 2014.

11. L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

12. R. T. Baillie and T. Bollerslev. Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics*, 52(1-2):91–113, 1992.

13. T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources? In *IJCNLP*, pages 356–364, 2013.

14. P. Ball. Counting google searches predicts market movements. *Nature*, 12879, 2013.

15. B. Batrinca and P. C. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116, 2015.

16. S. Bergsma, M. Dredze, B. Van Durme, T. Wilson, and D. Yarowsky. Broadly improving user classification via communication-based name and location clustering on twitter. In *HLT-NAACL*, pages 1010–1019, 2013.

17. T. M. Bernardo, A. Rajic, I. Young, K. Robiadek, M. T. Pham, and J. A. Funk. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research*, 15(7):e147, 2013.

18. J. Bian, U. Topaloglu, and F. Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM, 2012.

19. T. Bodnar and M. Salathé. Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 699–702. ACM, 2013.

20. B. Boecking, E. AU, and J. Schneider. Predicting events surrounding the egyptian revolution of 2011 using learning algorithms on micro blog data, 2014. Paper presented at the 2014 Internet, Politics, and Policy conference.

21. B. Boecking, M. Hall, and J. Schneider. Event prediction with learning algorithms: A study of events surrounding the egyptian revolution of 2011 on the basis of micro blog data. *Policy & Internet*, 7(2):159–184, 2015.

22. J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

23. E. Bothos, D. Apostolou, and G. Mentzas. Agent based information aggregation markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 449–454. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

24. U. S. C. Bureau. Population estimates, 2016.

25. J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

26. P. Burnap, W. Colombo, and J. Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 75–84. ACM, 2015.

27. M. P. Cameron, P. Barrett, and B. Stewardson. Can social media predict election results? evidence from new zealand. *Journal of Political Marketing*, pages 1–17, 2015.

28. S. Canada. Statistics canada: Canada's national statistical agency.

29. A. Ceron, L. Curini, S. M. Iacus, and G. Porro. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New Media & Society*, 16(2):340–358, 2014.

30. S. Chakravarty. Stealth-trading: Which traders' trades move stock prices? *Journal of Financial Economics*, 61(2):289–307, 2001.

31. S. Chancellor, Z. Lin, E. L. Goodman, S. Zerwas, and M. De Choudhury. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1171–1184. ACM, 2016.

32. S. Chancellor, T. Mitra, and M. De Choudhury. Recovery amid pro-anorexia: Analysis of recovery in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2111–2123. ACM, 2016.

33. H.-w. Chang, D. Lee, M. Eltaher, and J. Lee. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society, 2012.

34. L. E. Charles-Smith, T. L. Reynolds, M. A. Cameron, M. Conway, E. H. Lau, J. M. Olsen, J. A. Pavlin, M. Shigematsu, L. C. Streichert, K. J. Suda, et al. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one*, 10(10):e0139701, 2015.

35. C. Chen, W. Dongxing, H. Chunyan, and Y. Xiaojie. Exploiting social media for stock market prediction with factorization machine. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02*, pages 142–149. IEEE Computer Society, 2014.

36. J. Chen, H. Chen, D. Hu, J. Z. Pan, and Y. Zhou. Smog disaster forecasting using social web data and physical sensor data. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 991–998. IEEE, 2015.

37. J. Chen, E. Haber, R. Kang, G. Hsieh, and J. Mahmud. Making use of derived personality: The case of social media ad targeting. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2015.

38. L. Chen, W. Wang, and A. P. Sheth. Are twitter users equal in predicting elections? a study of user groups in predicting 2012 us republican presidential primaries. In *International Conference on Social Informatics*, pages 379–392. Springer, 2012.

39. X. Chen, Y. Cho, and S. Y. Jang. Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*, pages 63–68. IEEE, 2015.

40. Y. Chen, J. Zhao, X. Hu, X. Zhang, Z. Li, and T.-S. Chua. From interest to function: Location estimation in social media. In *AAAI*, 2013.

41. W.-Y. S. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse. Social media use in the united states: implications for health communication. *Journal of medical Internet research*, 11(4):e48, 2009.

42. E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

43. R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

44. M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. *ICWSM*, 133:89–96, 2011.

45. S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PloS one*, 6(8):e23610, 2011.

46. C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh. Using web and social media for influenza surveillance. In *Advances in Computational Biology*, pages 559–564. Springer New York, 2010.

47. C. D. Corley, C. Dowling, S. J. Rose, and T. McKenzie. Social sensor analytics: Measuring phenomenology at scale. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pages 61–66. IEEE, 2013.

48. D. Corney, C. Martin, and A. Göker. Spot the ball: Detecting sports events on twitter. In *European Conference on Information Retrieval*, pages 449–454. Springer, 2014.

49. A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.

50. A. Culotta. Estimating county health statistics with twitter. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1335–1344. ACM, 2014.

51. A. Culotta, N. R. Kumar, and J. Cutler. Predicting the demographics of twitter users from website traffic data. In *AAAI*, pages 72–78, 2015.

52. M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM, 2013.

53. M. De Choudhury, S. Counts, and E. Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.

54. M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638. ACM, 2014.

55. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, page 2, 2013.

56. M. De Choudhury, S. Jhaver, B. Sugar, and I. Weber. Social media participation in an activist movement for racial equality. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

57. M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. ACM, 2016.

58. M. De Choudhury, S. Sharma, and E. Kiciman. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1157–1170. ACM, 2016.

59. K. Denecke, P. Dolog, and P. Smrz. Making use of social media data in public health. In *Proceedings of the 21st International Conference on World Wide Web*, pages 243–246. ACM, 2012.

60. Q. Diao and J. Jiang. A unified model for topics, events and users on twitter. ACL, 2013.

61. N. Dokoohaki, F. Zikou, D. Gillblad, and M. Matskin. Predicting swedish elections with twitter: A case for stochastic link structure analysis. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1269–1276. ACM, 2015.

62. C. P. Dowling, J. J. Harrison, A. V. Sathanur, L. H. Sego, and C. D. Corley. Social sensor analytics: Making sense of network models in social media. In *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, pages 144–147. IEEE, 2015.

63. M. Duggan. Mobile messaging and social media 2015. pew research center; 2015, 2015.

64. M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden. Social media update 2014. *Pew Research Center*, 9, 2015.

65. N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

66. P. Earle. Earthquake twitter. *Nature Geoscience*, 3(4):221–222, 2010.

67. J. Eisenstein. What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369, 2013.

68. J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

69. E. Ellis. How the usgs uses twitter data to track earthquakes. *Twitter Data Stories, Twitter*, 2015.

70. Facebook. Facebook self-published usage statistics.

71. R. Feldman, O. Netzer, A. Peretz, and B. Rosenfeld. Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1779–1788. ACM, 2015.

72. K. Filippova. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488. Association for Computational Linguistics, 2012.

73. U. Franke and J. Brynielsson. Cyber situational awareness–a systematic review of the literature. *Computers & Security*, 46:18–31, 2014.

74. M. Gaurav, A. Srivastava, A. Kumar, and S. Miller. Leveraging candidate popularity on twitter to predict election outcome. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 7. ACM, 2013.

75. D. Gayo-Avello. Don't turn social media into another "literary digest" poll. *Communications of the ACM*, 54(10):121–128, 2011.

76. D. Gayo-Avello. I wanted to predict elections with twitter and all i got was this lousy paper—a balanced survey on election prediction using twitter data. *arXiv preprint arXiv:1204.6441*, 2012.

77. D. Gayo-Avello. No, you cannot predict elections with twitter. *IEEE Internet Computing*, 16(6):91–94, 2012.

78. D. Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, page 0894439313493979, 2013.

79. D. Gayo Avello, P. T. Metaxas, and E. Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011.

80. M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

81. F. Godin, J. Zuallaert, B. Vandersmissen, W. De Neve, and R. Van de Walle. Beating the bookmakers: Leveraging statistics and twitter microposts for predicting soccer results. In *KDD Workshop on Large-Scale Sports Analytics*, 2014.

82. P. Goldstein and J. Rainey. The 2010 elections: Twitter isn't a very reliable prediction tool. *Retrieved January*, 10:2012, 2010.

83. P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM, 2013.

84. B. J. Goode, S. Krishnan, M. Roan, and N. Ramakrishnan. Pricing a protest: Forecasting the dynamics of civil unrest activity in social media. *PloS one*, 10(10):e0139911, 2015.

85. F. J. Grajales III, S. Sheps, K. Ho, H. Novak-Lauscher, and G. Eysenbach. Social media: a review and tutorial of applications in medicine and health care. *Journal of medical Internet research*, 16(2):e13, 2014.

86. J. Grimmer. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(01):80–83, 2015.

87. D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.

88. K. Han, S. Lee, J. Y. Jang, Y. Jung, and D. Lee. Teens are from mars, adults are from venus: analyzing and predicting age groups with behavioral characteristics in instagram. In *Proceedings of the 8th ACM Conference on Web Science*, pages 35–44. ACM, 2016.

89. J. Harrison, E. Bell, C. Corley, C. Dowling, and A. Cowell. Assessment of user home location geoinference methods. In *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, pages 148–150. IEEE, 2015.

90. D. M. Hartley, N. P. Nelson, R. Arthur, P. Barboza, N. Collier, N. Lightfoot, J. Linge, E. Goot, A. Mawudeku, L. Madoff, et al. An overview of internet biosurveillance. *Clinical Microbiology and Infection*, 19(11):1006–1013, 2013.

91. D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

92. T. Hu, H. Xiao, J. Luo, and T.-v. T. Nguyen. What the language you tweet says about your occupation. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

93. A. Hürriyetoglu, N. Oostdijk, and A. van den Bosch. Estimating time to event from tweets using temporal expressions. In *Proceedings of the 5th Workshop on Language Analysis for Social Media*, pages 8–16. [Sl]: Association for Computational Linguistics, 2014.

94. B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

95. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

96. Y. Jia, X. Song, J. Zhou, L. Liu, L. Nie, and D. S. Rosenblum. Fusing social networks with deep learning for volunteerism tendency prediction. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

97. L. Jin, H. Takabi, and J. B. Joshi. Towards active detection of identity clone attacks on online social networks. In *Proceedings of the first ACM conference on Data and application security and privacy*, pages 27–38. ACM, 2011.

98. I. L. Johnson, S. Sengupta, J. Schöning, and B. Hecht. The geography and importance of localness in geotagged social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 515–526. ACM, 2016.

99. A. Jungherr, P. Jürgens, and H. Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t.o., sander, p.g., & welpe, i.m. "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social science computer review*, 30(2):229–234, 2012.

100. E. Kalampokis, E. Tambouris, and K. Tarabanis. Understanding the predictive power of social media. *Internet Research*, 23(5):544–559, 2013.

101. N. Kallus. Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 625–630. ACM, 2014.

102. A. Khatua, A. Khatua, K. Ghosh, and N. Chaki. Can# twitter_trends predict election results? evidence from 2014 indian general election. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 1676–1685. IEEE, 2015.

103. S. Kinsella, V. Murdock, and N. O'Hare. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.

104. F. Konkel. Tweets give usgs early warning on earthquakes. *The Business of Federal Technology*, 2013.

105. G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan. Combining heterogeneous data sources for civil unrest forecasting. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265. IEEE, 2015.

106. R. Krikorian. New tweets per second record, and how!, 2013.

107. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

108. A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.

109. V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72, 2012.

110. V. Lampos, D. Preotiuc-Pietro, and T. Cohn. A user-centric model of voting intention from social media. In *ACL (1)*, pages 993–1003, 2013.

111. D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

112. D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

113. K. Leetaru and P. A. Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2. Citeseer, 2013.

114. M. T. Lehrman, C. O. Alm, and R. A. Proano. Detecting distressed and non-distressed affect states in short forum texts. In *Proceedings of the Second Workshop on Language in Social Media*, pages 9–18. Association for Computational Linguistics, 2012.

115. J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

116. J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006.

117. E. Y. Li, C.-Y. Tung, and S.-H. Chang. The wisdom of crowds in action: Forecasting epidemic diseases with a web-based prediction market system. *International Journal of Medical Informatics*, 92:35–43, 2016.

118. L. Li, M. Sun, and Z. Liu. Discriminating gender on chinese microblog: A study of online behaviour, writing style and preferred vocabulary. *Screen*, 501:1197161814, 2014.

119. Q. Li, B. Zhou, and Q. Liu. Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion. In *Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on*, pages 359–364. IEEE, 2016.

120. Y. Li, J. Huang, and J. Luo. Using user generated online photos to estimate and monitor air pollution in major cities. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, page 79. ACM, 2015.

121. Y. Li, V. Rakesh, and C. K. Reddy. Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 247–256. ACM, 2016.

122. H. Lin, J. Jia, L. Nie, G. Shen, and T.-S. Chua. What does social media say about your stress? In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3775–3781, 2016.

123. A. W. Lo and A. C. MacKinlay. *A non-random walk down Wall Street*. Princeton University Press, 2002.

124. C.-T. Lu, S. Xie, X. Kong, and P. S. Yu. Inferring the impacts of social media on crowdfunding. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 573–582. ACM, 2014.

125. T. Mahmood and U. Afzal. Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In *Information assurance (ncia), 2013 2nd national conference on*, pages 129–134. IEEE, 2013.

126. M. Makrehchi, S. Shah, and W. Liao. Stock prediction using event-based sentiment analysis. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 337–342. IEEE, 2013.

127. B. G. Malkiel. The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, 17(1):59–82, 2003.

128. Y. Mao, W. Wei, B. Wang, and B. Liu. Correlating s&p 500 stocks with twitter data. In *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research*, pages 69–72. ACM, 2012.

129. M. Marchetti-Bowick and N. Chambers. Learning for microblogs with distant supervision: Political forecasting with twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612. Association for Computational Linguistics, 2012.

130. V. Martin. Predicting the french stock market using social media analysis. In *Semantic and Social Media Adaptation and Personalization (SMAP), 2013 8th International Workshop on*, pages 3–7. IEEE, 2013.

131. J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 459–468. ACM, 2013.

132. S. Mei, H. Li, J. Fan, X. Zhu, and C. R. Dyer. Inferring air pollution by sniffing social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 534–539. IEEE, 2014.

133. P. Meladianos, G. Nikolentzos, F. Rousseau, Y. Stavrakas, and M. Vazirgiannis. Degeneracy-based real-time sub-event detection in twitter stream. In *Ninth International AAAI Conference on Web and Social Media*, pages 248–257, 2015.

134. P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 165–171. IEEE, 2011.

135. S. E. Middleton, L. Middleton, and S. Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17, 2014.

136. G. Mishne, N. S. Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 155–158, 2006.

137. E. Mohammady Ardehaly and A. Culotta. Inferring latent attributes of twitter users with label regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–195, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

138. Moher, D and Liberati, A and Tetzlaff, J and Altman, DG and The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS Medicine*, 6(7), 2009.

139. S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4):e85, 2013.

140. E. L. Murnane and S. Counts. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1345–1354. ACM, 2014.

141. S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan. Planned protest modeling in news and social media. In *AAAI*, pages 3920–3927, 2015.

142. A. Nikfarjam and G. H. Gonzalez. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annu Symp Proc*, volume 2011, pages 1019–1026, 2011.

143. B. Obama. Presidential memorandum – climate change and national security. Technical report, Office of the Press Secretary, 2016.

144. B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

145. C. Oh and O. Sheng. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In *Proceedings of the Thirty Second International Conference on Information Systems*, 2011.

146. D. E. O'Leary. Twitter mining for discovery, prediction and causality: Applications and methodologies. *Intelligent Systems in Accounting, Finance and Management*, 22(3):227–247, 2015.

147. N. Oliveira, P. Cortez, and N. Areal. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 31. ACM, 2013.

148. D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput Biol*, 9(10):e1003256, 2013.

149. O. Ozdikis, P. Senkul, and H. Oguztuzun. Semantic expansion of hashtags for enhanced event detection in twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*. Citeseer, 2012.

150. E. Paiva. How social media strategy helps 'deadpool' to be a massive box office success. 2016.

151. M. J. Paul, M. Dredze, and D. Broniatowski. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*, 2014.

152. C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

153. J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

154. B. Perozzi and S. Skiena. Exact age prediction in social networks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 91–92. ACM, 2015.

155. A. Perrin, M. Duggan, L. Rainie, A. Smith, S. Greenwood, M. Porteus, and D. Page. Social media usage: 2005-2015. pew research center, 2015.

156. F. Pimenta, D. Obradovic, and A. Dengel. A comparative study of social media prediction potential in the 2012 us republican presidential preelections. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 226–232. IEEE, 2013.

157. A. Porshnev, I. Redkin, and A. Shevchenko. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 440–444. IEEE, 2013.

158. T. Preis, H. S. Moat, S. Bishop, P. Treleaven, and H. E. Stanley. Quantifying the digital traces of hurricane sandy on flickr. *Scientific Reports*, 3(3141), 2013.

159. D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717, 2015.

160. A. Press. Facebook reveals its trending topics are curated by humans, 2016.

161. V. Radosavljevic, M. Grbovic, N. Djuric, and N. Bhamidipati. Large-scale world cup 2014 outcome prediction based on tumblr posts. In *KDD Workshop on Large-Scale Sports Analytics*, 2014.

162. N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, et al. 'beating the news' with embers: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1799–1808. ACM, 2014.

163. T. Rao and S. Srivastava. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 119–123. IEEE Computer Society, 2012.

164. M. A. Razzaq, A. M. Qamar, and H. S. M. Bilal. Prediction and analysis of pakistan election 2013 based on sentiment analysis. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 700–703. IEEE, 2014.

165. P. Resnik, A. Garron, and R. Resnik. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural*, pages 1348–1353. Association for Computational Linguistics, 2013.

166. A. Ritter, E. Wright, W. Casey, and T. Mitchell. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*, pages 896–905. ACM, 2015.

167. D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

168. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

169. M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, et al. Digital epidemiology. *PLoS Comput Biol*, 8(7):e1002616, 2012.

170. M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

171. P. Saleiro, L. Gomes, and C. Soares. Sentiment aggregate functions for political opinion polling using microblog streams. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, pages 44–50. ACM, 2016.

172. E. T. K. Sang and J. Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60. Association for Computational Linguistics, 2012.

173. M. Sap, G. Park, J. C. Eichstaedt, M. L. Kern, D. Stillwell, M. Kosinski, L. H. Ungar, and H. A. Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1151. Association for Computational Linguistics, 2014.

174. W. S. Sarle. Stopped training and other remedies for overfitting. *Computing science and statistics*, pages 352–360, 1996.

175. M. Scharkow. The accuracy of self-reported internet use: A validation study using client log data. *Communication Methods and Measures*, 10(1):13–27, 2016.

176. H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, and M. Strohmaier. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

177. A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *ICWSM*, 2013.

178. I. Segura-Bedmar, R. Revert, and P. Martínez. Detecting drugs and adverse events from spanish health social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 106–115, 2014.

179. L. Shi, N. Agarwal, A. Agrawal, R. Garg, and J. Spoelstra. Predicting us primary elections with twitter. *URL: http://snap. stanford. edu/social2012/papers/shi. pdf*, 2012.

180. H.-H. Shuai, C.-Y. Shen, D.-N. Yang, Y.-F. Lan, W.-C. Lee, P. S. Yu, and M.-S. Chen. Mining online social data for detecting social network mental disorders. In *Proceedings of the 25th International Conference on World Wide Web*, pages 275–285. International World Wide Web Conferences Steering Committee, 2016.

181. M. Skoric, N. Poor, P. Achananuparp, E.-P. Lim, and J. Jiang. Tweets and votes: A study of the 2011 singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2583–2591. IEEE, 2012.

182. X. Song, Z.-Y. Ming, L. Nie, Y.-L. Zhao, and T.-S. Chua. Volunteerism tendency prediction via harvesting multiple social networks. *ACM Transactions on Information Systems (TOIS)*, 34(2):10, 2016.

183. N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

184. Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):1, 2015.

185. C. Tan, L. Lee, and B. Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*, 2014.

186. W.-H. Tang, M.-Y. Yeh, and A. J. Lee. Information diffusion among users on facebook fan pages over time: Its impact on movie box office. In *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*, pages 340–346. IEEE, 2014.

187. S. Thielman. Facebook fires trending team, and algorithm without humans goes crazy, 2016.

188. S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM, 2015.

189. J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014.

190. Z. Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*, 2014.

191. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

192. L. Vaughan and M. Thelwall. Search engine coverage bias: evidence and possible causes. *Information processing & management*, 40(4):693–707, 2004.

193. E. Velasco, T. Agheneza, K. Denecke, G. Kirchner, and T. Eckmanns. Social media and internet-based data in global systems for public health surveillance: A systematic review. *Milbank Quarterly*, 92(1):7–33, 2014.

194. S. Volkova and Y. Bachrach. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*, 2016.

195. S. Volkova, Y. Bachrach, and B. Van Durme. Mining user interests to predict perceived psycho-demographic traits on twitter. In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 36–43. IEEE, 2016.

196. S. Volkova, G. Coppersmith, and B. Van Durme. Inferring user political preferences from streaming communications. In *ACL (1)*, pages 186–196, 2014.

197. S. Volkova and B. Van Durme. Online bayesian models for personal analytics in social media. In *AAAI*, pages 2325–2331, 2015.

198. C. Wade. The reddit reckoning, 2014.

199. M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *European Conference on Information Retrieval*, pages 356–367. Springer, 2013.

200. M.-H. Wang and C.-L. Lei. Boosting election prediction accuracy by crowd wisdom on social forums. In *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 348–353. IEEE, 2016.

201. Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui. Regularity and conformity: location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284. ACM, 2015.

202. D. J. Watts, J. Peretti, and M. Frumin. *Viral marketing for the real world*. Harvard Business School Pub., 2007.

203. W. Wei, K. Joseph, W. Lo, and K. M. Carley. A bayesian graphical model to discover latent events from twitter. In *Ninth International AAAI Conference on Web and Social Media*, volume 1, 2015.

204. K. Weller. Accepting the challenges of social media research. *Online Information Review*, 39(3):281–289, 2015.

205. L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3, 2013.

206. C. B. Williams and G. Gulati. The political impact of facebook: Evidence from the 2006 midterm elections and 2008 nomination contest. *Politics and Technology Review*, 1(1):11–24, 2008.

207. T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

208. B. Wing and J. Baldridge. Hierarchical discriminative classification for text-based geolocation. In *EMNLP*, pages 336–348, 2014.

209. H.-H. Won, W. Myung, G.-Y. Song, W.-H. Lee, J.-W. Kim, B. J. Carroll, and D. K. Kim. Predicting national suicide numbers with social media data. *PloS one*, 8(4):e61809, 2013.

210. S.-H. Wu, M.-J. Chou, C.-H. Tseng, Y.-J. Lee, and K.-T. Chen. Detecting in-situ identity fraud on social network services: a case study on facebook. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 401–402. ACM, 2014.

211. D. Xu, P. Cui, W. Zhu, and S. Yang. Graph-based residence location inference for social media users. *IEEE MultiMedia*, 21(4):76–83, 2014.

212. J. Xu, T.-C. Lu, R. Compton, and D. Allen. Civil unrest prediction: A tumblr-based exploration. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 403–411. Springer, 2014.

213. S. Yardi and D. Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327, 2010.

214. A. Yates, N. Goharian, and O. Frieder. Extracting adverse drug reactions from social media. In *AAAI*, pages 2460–2467, 2015.

215. Q. You, S. Bhatia, T. Sun, and J. Luo. The eyes of the beholder: Gender prediction using images posted in online social networks. In *2014 IEEE International Conference on Data Mining Workshop*, pages 1026–1030. IEEE, 2014.

216. S. Yu and S. Kak. A survey of prediction using social media. *arXiv preprint arXiv:1203.1647*, 2012.

217. Y. Yu and X. Wang. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets. *Computers in Human Behavior*, 48:392–400, 2015.

218. D. Zeng, H. Chen, R. Lusch, and S.-H. Li. Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6):13–16, 2010.

219. C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han. Gmove: Group-level mobility modeling using geo-tagged social media. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

220. J. Zhang, X. Hu, Y. Zhang, and H. Liu. Your age is no secret: Inferring microbloggers' ages via content and interaction analysis. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

221. K. Zhang, Y.-R. Lin, and K. Pelechrinis. Eigentransitions with hypothesis testing: The anatomy of urban mobility. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

222. W. Zhang, S. Ram, M. Burkart, and Y. Pengetnze. Extracting signals from social media for chronic disease surveillance. In *Proceedings of the 6th International Conference on Digital Health Conference*, DH '16, pages 79–83, New York, NY, USA, 2016. ACM.

223. X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

224. S. Zhao, Y. Tong, X. Liu, and S. Tan. Correlating twitter with the stock market through non-gaussian svar. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, pages 257–264. IEEE, 2016.

225. W. X. Zhao, B. Shu, J. Jiang, Y. Song, H. Yan, and X. Li. Identifying event-related bursts via social media activities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1466–1477. Association for Computational Linguistics, 2012.

226. D. Zheng, T. Hu, Q. You, H. Kautz, and J. Luo. Inferring home location from user's photo collections based on visual content and mobility patterns. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, pages 21–26. ACM, 2014.

227. D. Zimbra, H. Chen, and R. F. Lusch. Stakeholder analyses of firm-related web forums: Applications in stock return prediction. *ACM Transactions on Management Information Systems (TMIS)*, 6(1):2, 2015.

228. B. Zou, V. Lampos, R. Gorton, and I. J. Cox. On infectious intestinal disease surveillance using social media content. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 157–161. ACM, 2016.