Predictive Analysis on Twitter: Techniques and Applications

Ugur Kursuncu[1,2], Manas Gaur[1], Usha Lokala[1], Krishnaprasad Thirunarayan[1], Amit Sheth[1], and I. Budak Arpinar[2]

[1]Kno.e.sis Center, Wright State University, Dayton, OH, USA
[2]Department of Computer Science, The University of Georgia, Athens, GA, USA
{ugur,manas,usha,prasad,amit}@knoesis.org
{kursuncu,budak}@uga.edu

Abstract. Predictive analysis of social media data has attracted considerable attention from the research community as well as the business world because of the essential and actionable information it can provide. Over the years, extensive experimentation and analysis for insights have been carried out using Twitter data in various domains such as healthcare, public health, politics, social sciences, and demographics. In this chapter, we discuss techniques, approaches and state-of-the-art applications of predictive analysis of Twitter data. Specifically, we present fine-grained analysis involving aspects such as sentiment, emotion, and the use of domain knowledge in the coarse-grained analysis of Twitter data for making decisions and taking actions, and relate a few success stories.

Keywords. Social media analysis, Citizen sensing, Community evolution, Event analysis, Sentiment-emotion-intent analysis, Spatio-temporal-thematic analysis, Election prediction, Harassment detection, Mental health, Demographic prediction, Drug trends, Stock Market prediction, Machine Learning, Semantic Social Computing.

1 Introduction

With the growing popularity of social media and networking platforms as an important communication and sharing media, they have significantly contributed to the decision making process in various domains. In the last decade, Twitter has become a significant source of user-generated data. The number of monthly active users was 330 million as of third quarter of 2017, and the number of daily active users was 157 million as of second quarter of 2017. Moreover, nearly 500 million tweets per day are shared on Twitter. Accordingly, significant technical advancements have been made to process and analyze social media data using techniques from different fields such as machine learning, natural language processing, statistics, and semantic web. This amalgamation and interplay of multiple techniques within a common framework have provided feature-rich analytical tools [1, 2], leading to valid, reliable and robust solutions.

Twitter provides multimodal data containing text, images, and videos, along with contextual and social metadata such as temporal and spatial information, and information about user connectivity and interactions. This rich user-generated data plays a significant role in gleaning aggregated signals from the content and making sense of public opinions and reactions to contemporary issues. Twitter data can be used for predictive analysis in many application areas, ranging from personal and social to public health and politics. Predictive analytics on Twitter data comprises a collection of techniques to extract information and patterns from data, and predict trends, future events, and actions based on the historical data.

Gaining insights and improving situational awareness on issues that matter to the public are challenging tasks, and social media can be harnessed for a better understanding of the pulse of the populace. Accordingly, state-of-the-art applications, such as Twitris [3] and OSoMe [2], have been developed to process and analyze big social media data in real time. Regarding availability and popularity, Twitter data is more common than data from web forums and Reddit[1]. It is a rich source of user behavior and opinions. Although analytical approaches have been developed to process Twitter data, a systematic framework to efficiently monitor and predict the outcome of events has not been presented. Such a framework should account for the granularity of the analysis over a variety of domains such as public health, social science, and politics, and it has been shown in Figure 1.

We discuss a predictive analysis paradigm for Twitter data considering prediction as a process based on different levels of granularity. This paradigm contains two levels of analysis: *fine-grained* and *coarse-grained.* We conduct fine-grained analysis to make tweet-level predictions on domain independent aspects such as sentiment, topics, and emotions. On the other hand, we perform coarse-grained analysis to predict the outcome of a real-world event, by aggregating and combining fine-grained predictions. In the case of fine-grained prediction, a predictive model is built by analyzing social media data, and prediction is made through the application of the model to previously unseen data. Aggregation and combination of these predictions are made from individual tweets form signals that can be used for coarse-grained predictive analysis. In essence, low-level signals from tweets, such as sentiment, emotions, volume, topics of interest, location and timeframe, are used to make high-level predictions regarding real-world events and issues.

In this chapter, we describe use of Twitter data for predictive analysis, with applications to several different domains. In Section 2, we discuss both processing and analytic techniques for handling Twitter data and provide details of feature extraction as well as machine learning algorithms. In Section 3, we explain a predictive analysis paradigm for Twitter that comprises two levels: fine-grained and coarse-grained. We also provide use cases, based on real-world events, of how coarse-grained predictions can be made by deriving more profound insights about a situation from social media using signals extracted through fine-grained

[1] https://goo.gl/Jo1h9U

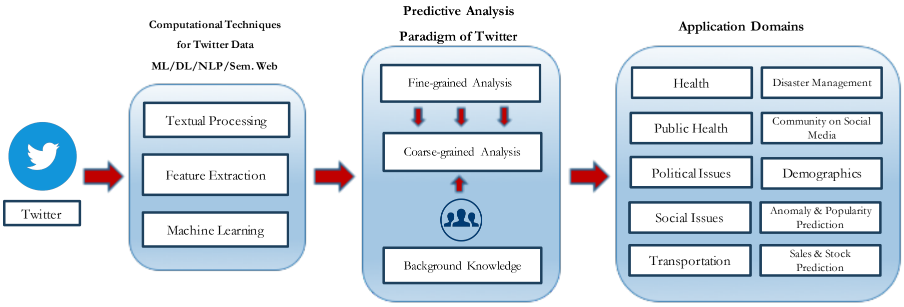

Fig. 1. Overview of Predictive Analysis on Twitter Data.

predictions. We also describe common domain-independent building blocks that can serve as the foundation for domain-specific predictive applications. In Section 4, we give further details on specific state-of-the-art applications of Twitter analytics that have been developed for different domains, such as public health, social and political issues. In Section 5, we conclude with a discussion of the impact of social media on the evolvement of real-world events and actions, challenges to overcome, for broader coverage and more reliable prediction. We also provide a comparative table relating techniques used with corresponding applications.

2 Language Understanding of Tweets

Novel processing and analysis techniques are required to understand and derive reliable insights to predict trends and future events from Twitter data due to their unique nature – it contains slangs, unconventional abbreviations and grammatical errors as a matter of course. Moreover, due to the evolving nature of many events, may it be political, sports, or disaster-related, collecting relevant information as the event unfolds is crucial [4, 5]. Overcoming the challenges posed by the volume, velocity, and variety of incoming social, big data is non-trivial [6]. Sole keyword-based crawling suffers from low precision as well as low recall. For instance, obtaining tweets related to marijuana legislation [7] using its street name spice pulls irrelevant content about pumpkin spice latte and spice in food. To improve recall without sacrificing precision, Sheth et al. [8] provided a solution for adapting and enhancing filtering strategies that (a) obtains customized tweet streams containing topics of user interest [9] by constructing a hierarchical knowledge base by analyzing each users tweets and profile information [10], (b) selects and employs a domain-specific knowledge graph (e.g., using the Drug

Abuse Ontology for opioid related analysis [11]) for focus, and (c) reuses a broad knowledge graph such as DBPedia for coverage and generality. In Twitter data analysis, the processing phase includes natural language processing using techniques such as TF-IDF, word2vec, stemming, lemmatization, eliminating words with a rare occurrence, and tokenizing. On the other hand, some of the commonly used techniques, such as removal of stop-words, have proven ineffective. Saif [12] has compared six different stop words identification methods over six different Twitter datasets using two well-known supervised machine learning methods and assessed the impact of removing stop words by observing fluctuations in the level of data sparsity, the size of the classifiers feature space and the classifier performance. Saif concludes that in most cases that removing stop words from tweets has a negative impact on the classification performance.

2.1 Unique Nature of Tweets

Twitters limit on the number of characters in a message encourages the use of unconventional abbreviations, misspellings, grammatical errors and slang terms. For instance, since a tweet was limited to 140 characters (until recent doubling to 280 character in December 2017), different sets of techniques and metadata have been considered to identify the best features to optimize the overall performance of the model being built. Due to the heterogeneous nature of the Twitter content, one can develop a variety of features [13] ranging from textual, linguistic, visual, semantic, network-oriented, to those based on the tweet and user metadata. Further, to handle tweets textual data, the extracted features, techniques and tools [3, 14, 15] have been customized to exploit as well as being robust concerning misspellings, abbreviations, and slangs. Gimpel et al. [14] addressed this problem in the context of part-of-speech (PoS) tagging, by developing a new tagset along with features specific to tweets, and reported 89% accuracy as opposed to Stanford tagger with 85% accuracy.

Tweets also include hashtags, URLs, emoticons, mentions, and emoji in their content. As these components contribute to the meaning of a tweet, it is imperative that we incorporate them in the analysis, on a par with textual content.

Hashtags are meant to help in categorizing tweet's topics. They are frequently used to collect and filter data as well as for sentiment [16–18], emotion [6], and topical analysis [19, 20]. Wang et al. [16] used hashtags in their topical hashtag level sentiment analysis incorporating co-occurrence and literal meaning of hashtags as features in a graph-based model and reported better results compared to a sentiment analysis approach at the tweet level. In emotion analysis, Wang et al. [6] collected about 2.5 million tweets that contain emotion-related hashtags such as #excited, #happy, and #annoyed, and used them as the self-labeled training set for developing a high accuracy, supervised emotion classifier.

URL presence in a tweet is usually indicative the content being an index for a longer explanatory story pointed to by the URL. Researchers found URLs in a tweet to be discriminative in various studies such as sentiment analysis [21, 22], popularity prediction [23, 24], spam detection [25]. They reported that the feature for URL presence in a tweet appeared as a top feature or has a substantial contribution to the accuracy of the model.

Emoticons (e.g., :), < 3) have been exploited by Liu et al. [26] in their Twitter sentiment analysis study, such as by interpreting :) as conveying positive sentiment and :(as conveying the negative sentiment. They used all tweets containing those emoticons as self-labeled training set and integrated them with the manually labeled training set [27]. They have achieved significant improvement over the model trained with only manually labeled data. Go et al. [21], and other researchers [28, 29] conducted sentiment analysis on Twitter in 2009, and they found that they were able to achieve a better accuracy using models trained with emoticon data.

Emoji is a pictorial representation of facial expressions, places, food and many other objects, being used very often on social media to express opinions and emotions on contemporary issues of contentions and discussions. The use of emoji is similar to emoticon since they both provide a shorter means of expression of an idea and thought. The difference is that an emoji use a small image for the representation as opposed to emoticon that uses a sequence of characters. Kelly et al. [30] studied the use of emoji in different contexts by conducting interviews and found that the use of emoji goes beyond the context that the designer intended. Novak et al. [31] created an emoji sentiment lexicon analyzing the sentiment properties of emojis, and they pointed that the emoji sentiment lexicon can be used along with the lexicon of sentiment-bearing words to train a sentiment classifier. On the other hand, Miller et al. [32] found that the emoji provided by different platforms are not interpreted similarly. Wijeratne et al. [33] gathered possible meanings of 2,389 emojis in a dataset called EmojiNet, providing a set of words (e.g., smile), its POS tag (e.g., verb), and its definition, that is called its sense. It associates 12,904 sense labels with 2,389 emojis, addressing the problem of platform-specific meanings by identifying 40 most confused emoji to a dataset.

2.2 Metadata for Tweet and User

There are mainly two types of metadata in a tweet object, namely, tweet metadata[2] and user metadata[3]. Tweet metadata contains temporal and spatial information along with user interactions and other information such as replies and language. On the other hand, user metadata contains information pertaining to the user that authored the tweet, such as screen-name and description. Some of the available metadata are described below.

2.2.1 Tweet Metadata

createdAt: This field contains the information on when the tweet was created, which is especially important when a time series analysis is being done [34].
favoriteCount: The users on Twitter can like a tweet, and this is one way of interacting with the platform. The number of likes for a tweet has been used as a feature in various applications that includes trend detection [34], identification

[2] TweetObject.https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

[3] UserObject.https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object

of influence and popularity.
inReplyToScreenName: If this field of the tweet object is not null, it is a reply to another tweet, and this field will hold the username of the user that authored the other tweet. This information is valuable, especially to predict the engagement of the audience over an issue that tweets relate to, and to find influential users.
geoLocation: the Twitter platform has a feature that can attach the users geolocation to the tweet, but this is up to the users to make it publically available. Most of the users prefer not to share their geolocation.
retweet_count: Twitter allows users to repost a tweet by retweeting to their audience, and the original tweet holds this field to keep how many times this tweet has been retweeted. This information is useful to incorporate the prediction of popularity and trending topics.

2.2.2 User Metadata

description: This field holds the description of the account. As this metadata carries information on characteristics of the user, it is mostly used in user classification.
followers_count: This field holds the number of followers the user has, and as it is changeable information over time, the information located in a specific tweet may not be up to date.
friends_count: Twitter calls the accounts that a user follows as "friends," but it is also known as "followees." The numbers of followers and followees are used to determine the popularity of user and topics.
statuses_count: Twitter also calls tweets as status, and in this case, status count refers to the number of tweets that a user has posted.

2.3 Network and Statistical Features

The users interact on the social networking platform Twitter with each other through follows, replies, retweets, likes, quotes, and mentions. Centrality metrics have been developed to compute and reveal users position and their importance based on their connections in their network. These centrality measures can help identify influential users. These metrics include in-degree, out-degree, closeness, betweenness, PageRank and eigenvector centrality. Closeness centrality is defined by Freeman [35] as the sum of distances from all other nodes, where the distance from a node to another is defined as the length (in links) of the shortest path from one to the other. The smaller the closeness centrality value, the more central the node. Betweenness [36] measures the connectivity of a node by computing the number of shortest paths which pass through the node. This aspect makes this node, a user in a Twitter social network, an essential part of the network as it controls the flow of information in the network. Therefore, removing this node would disconnect the network. EigenVector [37, 38] metric measures the importance of a node based on the importance of its connections within the network. Therefore, the more critical connections a node gets, the more critical the node becomes. These metrics were used in a user classification application as features by Wagner et al. [15] because of the intuition that similar users would have similar network connectivity characteristics.

Statistical features such as min, max, median, mean, average, standard deviation, skewness, kurtosis, and entropy can be computed for several data attributes [34]. Machine learning determines a subset of these features that have the discriminative power necessary for particular applications and domains, especially for predicting user behaviors and user types [39]. For instance, [34] extracted statistical features of a user, tweet, network. The statistical analysis was done over attributes such as senders follower count, originators followee count, the time between two consecutive tweets, and the number of hashtags in a tweet. They conducted a time series analysis to predict if a trending meme is organic or promoted by a group. On the other hand, [39] utilized statistical features to predict the type of users on social media based on their political leanings, ethnicity, and affinity for a particular business. As they classified users, they computed statistical characteristics of tweeting behavior of users such as average number of messages per day, average number of hashtags and URLs per tweet, average number and standard deviation of tweets per day.

2.4 Machine Learning and Word Embeddings

Machine learning algorithms play a crucial role in the predictive analysis for modeling relationships between features. It is well-known that there is no universal optimal algorithm for classification or regression task, and in fact requires us to tailor the algorithm to the structure of the data and the domain of discourse. Recent survey papers [40–43] and our comparative analysis (see Table 1) of related influential studies show what algorithms we found to perform well for various applications. As can be seen, this covers a wide variety – Random Forest, Naive Bayes, Support Vector Machine, Artificial Neural Networks, ARIMA and Logistic Regression.

Furthermore, deep learning (a.k.a advanced machine learning) enhanced the performance of learning applications. Deep learning is a strategy to minimize the human effort without compromising performance. It is because of the ability of deep neural networks to learn complex representations from data at each layer, where it mimics learning in the brain by abstraction[4]. The presence of big data, GPU, and sufficiently large labeled/unlabeled datasets improve its efficacy. We discuss some of the applications that make use of deep learning for prediction task on social media in section 4.

Textual data processing benefits from the lexico-semantic representation of content. TF-IDF [44], Latent/Hierarchical Dirichlet Allocation(LDA/HDA) [45], Latent Semantic Analysis (LSA) [46] and Latent Semantic Indexing have been utilized in prior studies for deriving textual feature representations. In a recent paper [47], they put forward a word embedding approach called Word2Vec that generates a numerical vector representation of a word that captures its contextual meaning incorporating its nearby words in a sentence. Training the word embedding model on a problem-specific corpus is essential for high-quality

[4] How do Neural networks mimic the human brain? https://www.marshall.usc.edu/blog/how-do-neural-networks-mimic-human-brain

domain-specific applications, since the neighborhood set of words for an input term impacts its word embedding. For instance, pre-trained models of word2vec on news corpora generate poor word embeddings over a Twitter corpus. Wijeratne et al. [48] used word embeddings to further enhance the prediction of gang members on Twitter by training their model on a problem-specific corpus.

2.5 Multi-modality on Twitter

Visual elements such as images and videos are often used on social media platforms. While users can attach images and videos to their tweets, they can also upload a profile image and a header image. Since the latter images are mostly related to the users characteristics, personality, interest or a personal preference, these images are mostly used for classification of account type (e.g., media, celebrity, company), detection of user groups [49, 48] and identification of demographic characteristics (e.g., gender, age) [50]. Balasuriya et al. [49] used the profile image of users in their feature set for finding street gang members on Twitter since gang members usually set their profile image in a particular way to intimidate other people and members of rival gangs. They retrieved a set of 20 words and phrases for each image through the Clarifai[5] web service to be used as features. As image processing is costly regarding time and computational resources required for training a model to retrieve information from images, it is usually preferred to use off-the-shelf web services that provide cheaper, yet effective alternative, for scalable social media analytics.

3 Prediction on Twitter Data

Gaining understanding about and predicting an events outcome and its evolvement over time using social media, requires incorporation of analysis of data that may differ in granularity and variety. As tools [14, 51] are developed and customized for Twitter, its dynamic environment requires human involvement in many aspects. For instance, verification of a classification process [52] and annotation of a training dataset [33, 53, 54] are essential in the predictive analysis that can benefit from human expert guidance in creating ground truth dataset. Social media analysis in the context of complex and dynamic domains [55–57] is challenging. Our approach to overcoming this challenge and dealing with a variety of domains is to customize domain independent building blocks to derive low-level/fine-grained signals from individual tweets. Then, we aggregate and combine these signals to predict high-level/coarse-grained domain-specific outcomes and actions with a human in the loop.

3.1 A Predictive Analysis Paradigm for Twitter

We consider predictive analysis on Twitter data as a two-phase approach: The first phase is fine-grained predictive analysis and the second phase is coarse-grained predictive analysis. An illustration of this paradigm is depicted in figure 2. The fine-grained analysis is a tweet-level prediction for individual signals, such as sentiment and emotions, about an event that is being monitored. This low-level prediction is made by building a predictive model that employs feature

[5] https://www.clarifai.com

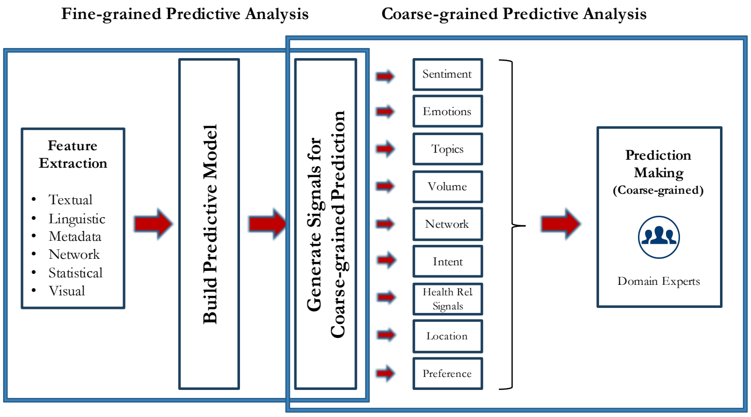

Fig. 2. Two level Predictive Analysis Paradigm for Twitter.

engineering and machine learning algorithms. Aggregating the tweet-level predictions for a specific time frame and location generates signals. For instance, a predictive model for sentiment predicts the sentiment of each tweet about an event in question as negative (-1) neutral (0) or positive (+1), and we produce a signal between -1 and +1 for a particular location and time frame. A collection of such signals (e.g., emotions, topics) helps domain experts form insights while monitoring or predicting the outcome of an event, in their higher level analysis. Extraction of these signals is discussed further in subsequent section.

Coarse-grained analysis is a higher level prediction involving outcomes and trends of a real-world event, such as elections[58], social movements[59] and disaster coordination[60–63]. In this case, we gather the signals which we generated from the fine-grained predictions and make a judgment call for the outcome by making sense of these signals in the context of the event and the related domain. Sentiment, emotions, volume, topics, and interactions between Twitter users can be considered as signals, while the importance and informativeness of each of these parameters may vary depending on the event and its domain. For instance, gauging the sentiment of a populace towards an electoral candidate would be very significant to predict the outcome of an election [56], but the same kind of information may not be as critical in the context of disaster management because, in the latter case, the sentiment may be largely negative. Further, for reliable decision making, the sentiment must be interpreted in a broader context. Predominantly positive sentiment towards democratic candidates in California is not as significant as that in Ohio. Similarly, the context provided by county demographics may be crucial in generalizing, predicting, and determining the outcome of an election. Moreover, temporal and spatial context plays an important role to understand the ongoing events better and obtain more profound insights. In US presidential elections, some states, called the swing states (as the electorates choice has changed between Republican and Democratic candidates through the previous elections in these states), typically determine the eventual outcome of US elections. Therefore, narrowing down the analysis to the state level and gathering signals from these particular states would meaningfully contribute to the prediction of the outcome of the Presidential election and the future direction of the country.

In general, prediction analytics requires domain-specific labeled datasets created with the assistance of domain experts, and customization of feature space, classification algorithm, and evaluation. Real world events have a dynamic nature in which critical happenings may change the course of discussions on social media. For example, breaking news about a candidate in an election may change the vibe in echo chambers of Twitter; thus, affecting the public opinion in one or another direction. For this reason, it is imperative to conduct the analysis accounting for essential milestone events happening during the process. Therefore, the analysis of such events would require an active learning paradigm that incorporates evolving domain knowledge in real-time.

3.2 Use Cases for Coarse-grained Prediction

Coarse-grained prediction requires taking into account many signals, and evaluating them concerning both present and historical context that varies with location and time frame. Importance of the signals in some domains and their related events may vary, and sole use of these signals would not be sufficient to make a reliable judgment call, although these signals are essential parameters in a real-world event context. For instance, an election usually whips up discussions on various sub-topics, such as unemployment, foreign policy; and necessitates proper cultivation of a diverse variety of signals following contextual knowledge of the domain [56]. We provide two use cases in this subsection to illustrate how a coarse-grained or high-level predictive analysis can be conducted.

3.2.1 US 2016 Presidential Election

During the 2016 US Presidential elections where swing states played a key role in determining the outcome, many polling agencies failed to predict it accurately[6][7]. On the other hand, researchers[8] conducted a real-time predictive analysis using a social media analytics platform [3], making the prediction accurately before the official outcome was announced, by analyzing the state-level signals, such as from Florida and Ohio. Temporal aspect was also important in this use case to explain the evolution of the public opinion based on milestone events over the period of the election, as well as the election day because people tend to express, who they voted in the same day. They analyzed 60 million tweets by looking at the sentiment, emotions, volume, and topics narrowing down their analysis to state-level. On the election day, they focused on specific states such as Florida,

[6] http://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/

[7] https://goo.gl/mFtzvb

[8] https://goo.gl/AJVpKf

which, before the election day, they predicted would be a pathway for Donald Trump to win the election[9]. In their analysis of Florida, volume and positive emotion (joy) for Trump was higher, whereas positive sentiment for Clinton was higher, eliciting report[10] such as limited to professed votes from Florida until 1pm is not looking in her favor. Later in the day, the volume of tweets for Trump increased to 75% of all tweets based on the hashtag "#ivoted". Particularly in critical states of Florida, North Carolina, and Michigan, volume and positive emotions for Trump were significantly higher than for Clinton, although the sentiment was countering the overall signal. They made the call that the winner of Presidency and Congress as Donald Trump and the GOP respectively. While conducting this analysis [56], they noticed that the predictive model that they have built for sentiment signal was not successful due to the dynamic nature of the election with changing topics in conversations. A similar analysis was made for UK Brexit polls in 2012 by the same researchers, correctly predicting the outcome utilizing the volume and sentiment signals[11][12][13].

3.2.2 US Gun Reform Debate 2018

Researchers[14] monitored gun reform discussions on Twitter to predict the public support using the Twitris platform after the tragic shooting at a high school in Parkland, Florida, in February 2018. The public started demanding a gun control policy reform, and it has attracted the attention of legislative and executive branches of both state and federal governments. As polls measured the public opinion[15], researchers reported that the public support for a gun reform on social media was increasing over time since the Parkland shooting, confirming the overall outcome of these polls. They observed that reactions from public on social media in terms of the volume, sentiment, emotions and topics of interest, are strongly aligned with the milestone events related to this issue such as (i) POTUS (President of the United States) meeting with families of the victims on February 21, (ii) CPAC(Conservative Political Action Conference) between February 22 and 24, (iii) POTUS meeting with lawmakers on February 28 expressing strong support for a gun control policy change. These events significantly affected the public opinion on social media based on the aforementioned signals.

At the beginning of the gun reform discussions on social media, sentiment for pro-gun reform tweets was strong whereas the sentiment for anti-gun reform was relatively weak. However, the CPAC meeting changed the climate on social media, and it significantly boosted the momentum of anti-gun reform tweets, especially after the NRA (National Rifle Association) CEO Wayne LaPierres speech in the morning of February 22[16]. Overall the volume of tweets for pro-gun

[9] https://goo.gl/sh7WNr

[10] https://goo.gl/iCqzk3

[11] https://goo.gl/i2Ztm6

[12] https://goo.gl/dFCGL9

[13] https://goo.gl/2EhSma

[14] http://blog.knoesis.org/2018/04/debate-on-social-media-for-gun-policy.html

[15] http://time.com/5180006/gun-control-support-has-surged-to-its-highest-level-in-25-years/

[16] https://goo.gl/kgbqWC

reform was mostly higher than the anti-gun overhaul, except between February 22 and February 24[17], which covers the CPAC meeting where NRA CEO, VP Pence, and POTUS gave speeches. It surged the volume, positive sentiment and emotions in anti-gun reform posts radically, and those parameters for pro-gun reform posts dropped in the same manner. Effect of the meeting lasted a few days, and boycott calls for NRA and NRAs sponsors started to pick up in the meantime. After the meeting, sentiment for pro-gun reform tweets increased consistently, and the emotions expressed in pro-gun reform tweets became more intensified.

Emotions in anti-gun reform tweets were intense especially during and after the CPAC meeting, but later emotions in pro-gun reform tweets took over. Especially volume, positive sentiment, and emotions were overwhelmingly high right after the POTUS meeting with lawmakers on Wednesday, February 28, expressing his support for a gun policy reform.

Furthermore, some of the most popular topics that users were discussing in their tweets included midterm elections, parkland students, boycott the nra, stupid idea and individual freedoms, where pro-gun reform arguments were expressed more frequently. The topic of midterm elections being one of the most popular topics on social media in gun reform discussions, also suggests that politicians from both Democrats and Republicans sensed the likely effect of this public opinion change on the midterm elections on November 2018. They have concluded in their predictive analysis that the public support for gun reform was significantly higher based on the signals they observed in the context of related events.

3.3 Extraction of Signals

We make predictions for the outcome of real-world events based on the insights we collect from big social data, and these insights are extracted as various signals such as sentiment, emotions, volume, and topics. The sentiment is a qualitative summary of opinions on a particular issue, and sentiment analysis techniques are utilized to extract such information computationally. The emotional analysis provides another stream of qualitative summary that is expressed by users about a particular event. The volume of tweets is an important signal about the engagement of the public in an event or an issue of consequence. Topical analysis is a process that extracts topics that contain particular themes in the domain of interest. We can produce and make use of more specific signals depending on the domain such as preference, intent, and symptoms. The signals described below are commonly used parameters in higher level prediction tasks, and we describe related state-of-the-art applications and their technical details in the following.

3.3.1 Sentiment Analysis

Sentiment is one of the essential signals that can be used to measure the public opinion about an issue. As users on Twitter express their opinions freely, sentiment analysis of tweets attracted the attention of many researchers. Their ap-

[17] `https://goo.gl/LMFu3B`

proaches differ regarding the feature set, machine learning algorithm, and text processing techniques. Considering feature set, [18] used n-grams, POS-tags, emoticons, hashtags and subjectivity lexicon for sentiment analysis. For machine learning, Naive Bayes, SVM, and Conditional Random Fields (CRF) have been employed, and Naive Bayes has shown good performance [29]. Also, text processing techniques like stopwords removal, word-pattern identification, and punctuation removal have shown to improve sentiment analysis in [17]. Nguyen et al. [64] used time series analysis to be able to predict the public opinion so that the stakeholders on a stock market can react or pro-act against the public opinion by posting balanced messages to revert the public opinion based on the measurement that they performed using social media. Their objective was to use the sentiment change over time by identifying key features that contribute to this change. They measured the sentiment change regarding the fraction of positive tweets. They employed SVM, logistic regression and decision tree, and found that SVM and logistic regression provided similar results outperforming the decision tree. They modeled the sentiment change overall twitter data and achieved around 73% F-score on sentiment prediction using time series analysis. [65] employs a deep learning approach combining convolutional and gated recurrent neural network (CGRNN) for a diverse representation of tweets for sentiment analysis. Such a system was trained on GloVe word embedding created on a crawled dataset. The system was ranked among the top 10, evaluated using average F1 score, average recall, mean absolute error (MAE), Kullback-Leibler divergence (KLD), and EMD score [66] for SemEval-2016 sub-tasks B, C, D, E. Exclusion of hand-crafted features and improved performance on SemEval 2016 shows the potency of the approach.

3.3.2 Emotion Analysis

Identification of emotions in tweets can provide valuable information about the public opinion on an issue. Wang et al. [6] predicted seven categorical emotions from the content of tweets using 131 emotion hashtags and utilizing the features such as n-grams, emotion lexicon words, part-of-speech tags, and n-gram positions. They used two machine learning algorithms: LIBLINEAR and Multinomial Naive Bayes. In a similar study, Lewenberg et al. [54] examined the relationship between the emotions that users express and their perceived areas of interest, based on a sample of users. They used Ekmans emotion categories and crowdsourced the task of examining the users and their tweets content to determine the emotions as well as their interest areas. They created a tweet-emotion dataset consisting of over 50,000 labeled tweet-emotion pairs, then trained a logistic regression model to classify the emotions in tweets according to emotion categories, using textual, linguistic and tweet metadata features. The model predicted a users emotion score for each emotion category, and they determined the user's interest in areas such as sports, movies, technical computing, politics, news, economics, science, arts, health, and religion.

3.3.3 Topical Analysis

Topical analysis is one of the essential strategies under the umbrella of information extraction techniques that capture semantically relevant topics from the social media content [67]. Extraction of topics in the context of social media analysis helps understand the subtopics associated with an event or issue and what aspects of the issue have attracted the most attention from the public. As discussed in use cases for elections and gun reform debate, it is imperative to have the topics extracted from tweets for a better understanding of the underlying dynamics of relevant discussions. Chen et al. [68] associated topics of interest with their relative sentiment to monitor the change in sentiments on the extracted topics. Furthermore, utilizing the extracted topics as features for a supervised model improved the performance of the classification task in [69]. In [70], researchers assessed quality of topics using coherence analysis, context-sensitive topical PageRank based ranking and probabilistic scoring function. This approach was used in a crime prediction application [71].

3.3.4 Engagement Analysis

The volume is the size of the dataset that has been collected and indicates the user engagement on an event being monitored. In general, the larger the dataset, the better is the accuracy and consistency of a predictive model because it minimizes the possibility of bias. Engagement analysis enables human experts to improve their confidence in the learned representations/patterns for an accurate high-level prediction. However, while maintaining the sufficient size of the dataset to make reliable predictions from representative data is critical, data collection strategies need to be chosen strategically since relying solely on keyword-based crawling can bring in noise and irrelevant[72] data from a different context into the dataset. Therefore, a suitable filtering mechanism is essential for better quality data with high recall as well as precision. A semantic filtering mechanism [8, 73] as in the Twitris platform, can be implemented that selects and employs a domain-specific knowledge graph (e.g., using the Drug Abuse Ontology for related opioid analysis [11]) for precision, and reuses a broad knowledge graph such as DBPedia for coverage and generality (see section 2). Thus, a significant and relevant dataset can be collected with high recall and precision that will allow one to obtain insights on the user engagement.

4 Applications

Twitter data has enabled researchers and analysts to deal with diverse domains ranging from healthcare, finance, and economy to socio-political issues and crisis management. Approaches to retrieve as much information as possible requires the inclusion of domain-specific features as well as the use of domain knowledge in the analysis. In this section, we provide a list of domains where predictive analysis applications on Twitter were implemented, along with the technical details. A comprehensive table is also included at the end to give a comparative overview of application domains, the features and machine learning algorithms being used, and their performance. The included applications were selected because they were state-of-the-art in their respective domains or had been influential. The applications that we describe in this section combine a variety of signals that can be the basis for coarse-grained predictive analysis. Since some of the applications in this section make use of the Twitris platform; therefore, we first provide background information about the platform. Purohit et al. [1] introduced the Twitris platform for citizen sensing that performs analysis of tweets, complemented by shared information from contextually relevant Web of Data and background knowledge. They describe it as a scalable and interactive platform which continuously collects, integrates, and analyzes tweets to give more profound insights. They demonstrate the capabilities of the platform with an analysis in various dimensions including spatio-temporal-thematic, people-content network, and sentiment-emotion-subjectivity, with examples from business intelligence including brand tracking, advertising campaigns, social/political unrests, and disaster events.

4.1 Healthcare

Twitter data can be employed to shed light on many healthcare and disease-related aspects of contemporary interest, ranging from Alzheimer and dementia progression [74] to eating disorders [75] and mental health problems [76, 77]. We focus on applications to glean depression in individuals or at a community level using self-reports about these conditions, their consequences, and patient experiences on Twitter.

Depression is a condition that a sizable population in all walks of life experiences in their daily life. Social media platforms including Twitter has been used to voluntarily express the mood changes and feelings as they arise. From these tweets, it is possible to predict whether a user is depressed or not, what symptoms they show as well as the reasons for their depressive mood. Some examples indicative of depression as expressed in tweets[18] include: "I live such a pathetic life.", "Cross the line if you feel insecure about every aspect of your life." ,"That's how depression hits. You wake up one morning afraid that you're going to live.", and "Secretly having a mental breakdown because nothing is going right and all motivation is lost.". These tweets epitomize the expression of emotional tumult that may underlie subsequent conscious actions in the physical world.

An interesting study by Yazdavar et al. [76] explored the detection of clinical depression from tweets by mimicking the PHQ-9 questionnaire which clinicians administer to detect depression in patients. This study is different from traditional clinical studies that use questionnaires and self-reported surveys. They crawled 23M tweets over 45K twitter users to uncover nine significant depressive symptoms; (1) Lack of Interest, (2) Feeling Down, (3) Sleep Disorder, (4) Lack of Energy, (5) Eating Disorder, (6) Low Self-esteem, (7) Concentration Problems, (8) Hyper/Lower Activity, and (9) Suicidal Thoughts. A probabilistic topic model with a semi-supervised approach is developed to assess clinical depression symptoms. This hybrid approach is semi-supervised in that it exploits a

[18] These tweets were modified before we share them in this chapter.

lexicon of depression symptoms as background information (top-down) and combines it with generative model gleaned from the social media data (bottom-up) to achieve a precision of 72% on unstructured text.

De Choudhury et al. [53] predicted the depression in an individual by exploiting their tweets. For ground truth dataset, they used, crowdsourcing to collect and label data. They utilized tweet metadata, network, statistical, textual and linguistic features, and time series analysis over a year of data to train an SVM model, obtaining an accuracy of 0.72.

The extraction of the location of people who experience depression using textual and network features can further assist in locating depression help centers. [78] utilizes a multiview[19] and deep learning based model, to predict the user location. The multi-entry neural network architecture (MENET) developed for location prediction uses words, the semantics of the paragraph (using doc2vec [79]), network features and topology (using node2vec [80]) and time-stamps to deduce users location. They achieved an accuracy over 60% for GeoText[20], UT-Geo11[21] and 55% for TwitterWorld[81]. Furthermore, MENET achieves an accuracy of 76% in region classification and 64.4% in state classification using GeoText dataset.

4.2 Public Health

Social media platforms including web forums, Reddit and Twitter, has become a venue where people seek advice and provide feedback for problems concerning public health. These conversations can be leveraged to predict trends in health-related issues that may threaten the well-being of the society. Moreover, caregivers have also seen these sources to be a game changer in its potential for actionable insights because of the information circulation. Particularly, cannabis legalization issue in the U.S. has been a trending topic[22] in the country as well as social media. Prior research on Twitter data analysis in this domain proved that it is an essential tool for epidemiological prediction of emerging trends.

Existing studies have involved identifying syntactic and statistical features for public health informatics, such as PREDOSE (PRescription Drug abuse Online Surveillance and Epidemiology) which is a semantic web platform that uses the web of data, background domain knowledge and manually created drug abuse ontology for extraction of contextual information from unstructured social media content. PREDOSE performs lexical, pattern recognition (e.g., slang term identification), trend analysis, triple extraction (subject-predicate-object) and content analysis. It is helpful in detecting substance abuse involving marijuana and related products. Not only can it analyze generic marijuana but also its concentrates like butane hash oil, dabs, and earwax that are used in the form of vaporizers or inhalers. In a similar analysis of Twitter data, the marijuana concentrate use and its trends were identified in states where cannabis was le-

[19] http://www.wcci2016.org/document/tutorials/ijcnn8.pdf

[20] https://www.cs.cmu.edu/~ark/GeoText/README.txt

[21] http://www.cs.utexas.edu/~roller/research/kd/corpus/README.txt

[22] http://www.pewresearch.org/fact-tank/2018/01/05/americans-support-marijuana-legalization/ft_18-01-05_marijuana_line_update/

galized as well as not legalized. In 2014, utilizing the eDrugTrends[23] Twitris platform, researchers collected a total of 125,255 tweets for a two-month period, and 22% of these tweets have state-level location information[82]. They found that the percentage of dabs-related tweets was highest in states that allowed recreational or medicinal cannabis use and lowest in states that have not passed medical cannabis laws, where the differences were statistically significant. A similar study in 2015 [83] reported adverse effects of Cannabis edibles and estimated the relationship between edibles-related tweeting activity and local cannabis legislation. Another study [82] was to automatically classify drug-related tweets by user type and the source of communication as to what type of user has authored the tweet, where the user types are defined as user, retailer and media. They employed supervised machine learning techniques incorporating the sentiment of tweets (e.g., positive, negative, neutral).

4.3 Political Issues

Political discussions on Twitter, which capture dynamic evolvement of public opinion, can directly impact the outcome of any political process. Arab Spring demonstrations [84–86] in the middle eastern countries, Gezi protests [87, 85] in Turkey, as well as US Presidential elections in 2016 involving influence peddling on several social media platforms [88] provide impactful illustrative examples. Researchers have explored user classification and profiling in the context of such political events on Twitter to predict the issue trends and eventual outcome.

Researchers [39, 89, 90] used Twitter data to predict political opinions of users based on linguistic characteristics (e.g., Tf-IDF) of their tweet content. While classification of users based on their political stance on Twitter has been well studied, Cohen et al. [91] have claimed that much of the studies and their datasets to date have covered very narrow portion of the Twittersphere, and their approaches were not transferable to other datasets. Pennacchiotti et al. [39] focused on the user profiling task on Twitter, and used user-centric features such as profile, linguistic, behavioral, social and statistical information, to detect their political leanings, ethnicity, and affinity for a particular business.

Moreover, prediction of dynamic groups of users has been employed [58] to monitor the polarity during a political event by analyzing tweeting behavior and content through clustering. Usage of hashtags and URL, retweeting behaviors and semantic associations between different events were key to clustering. 56% of the Twitter users participated in 2012 US Republican Primaries by posting at least one tweet, while 8% of the users tweeted more than 10 tweets. 35% of all users mostly retweet, separating them from the remaining. In terms of dynamic user groups, they formed the following bilateral groups: silent majority and vocal minority, high and low engaged users, right and left-leaning users, where users were from different political beliefs and ages. They analyzed these dynamic groups of users to predict the election outcomes of Super Tuesday primaries in 10 states. They also reported that the characterization of users by tweet properties (frequency of engagement, tweet mode, and type of content) and

[23] `http://wiki.knoesis.org/index.php/EDrugTrends`

political preference provided insights to make reliable predictions. 8 weeks of data comprising 6,008,062 tweets from 933,343 users about 4 Republican candidates: Newt Gingrich, Ron Paul, Mitt Romney and Rick Santorum, was analyzed to assess the accuracy of predicting the winner. Prediction of user location using a knowledge base such as LinkedGeoData[24] in tweets also contributed to the election prediction. Furthermore, an error of 0.1 between the prediction and actual votes attest to the efficacy of the approach. Such a low error rate in prediction is attributed to original tweets (not retweets) from users who are highly engaged and right leaned.

4.4 Social Issues

Social issues and related events have been a part of discussions on Twitter, which gives opportunities to the researchers to address problems concerning individuals as well as the society at large. Solutions to such problems can be provided by measuring public opinion and identification of cues for detrimental behavior on Twitter by employing predictive analysis. We explain three problems and their respective solutions in this subsection.

4.4.1 Harassment

Harassment[25] is defined as an act of bullying an individual through aggressive offensive word exchanges leading to emotional distress, withdrawal from social media and then life. According to a survey from Pew Research Center[26], 73% of the adult internet users have observed, and 40% have experienced harassment, where 66% percent of them are attributed to social media platforms. Also, according to a report from Cyberbullying[27] research center, 25% of teenagers claimed to be humiliated online. While it is imperative to solve this problem, frequency and severe repercussions of online harassment exhibit social and technological challenges.

Prior work [92] has modeled harassment on social media to identify the harassing content which was a binary classification approach. However, in their predictive analysis, the context, network of users and dynamically evolving communities shed more light on the activity than pure content-based analysis. For instance, sarcastic communication between two friends on social media may not be conceived as harassment while the aggressive conversation between two strangers can be considered as an example of bullying. For reliably identifying and predicting harassment on Twitter, it is essential to detect language-oriented features (e.g., negation, offensive words), emotions, and intent. [93] employs machine learning algorithms along with word embedding, and DBpedia knowledge graph to capture the context of the tweets and user profiles for harassment prediction.

Edupuganti et al. [94] focused on reliable detection of harassment on Twitter by better understanding the context in which a pair of users is exchanging messages, thereby improving precision. Specifically, it uses a comprehensive set

[24] http://linkedgeodata.org/About

[25] http://wiki.knoesis.org/index.php/Context-Aware_Harassment_Detection_on_Social_Media

[26] http://www.pewinternet.org/2014/10/22/online-harassment/

[27] http://cyberbullying.us/facts

of features involving content, profiles of users exchanging messages, and the sequence of messages, we call conversation. By analyzing the conversation between users and features such as change of behavior during their conversation, length of conversation and frequency of curse words, the harassment prediction can be significantly improved over merely using content features and user profile information. Experimental results demonstrate that the comprehensive set of features used in our supervised machine learning classifier achieves F-score of 88.2 and Receiver Operating Characteristic (ROC) of 94.3. Kandakatla et al. [95] presents a system that identifies offensive videos on YouTube by characterizing features that can be used to predict offensive videos efficiently and reliably. It exploits using content and metadata available for each YouTube video such as comments, title, description, and the number of views to develop Nave Bayes and Support Vector Machine classifiers.The training dataset of 300 videos and test dataset of 86 videos were collected, and the classifier obtained an F-Score of 0.86.

4.4.2 Gang Communities & Their Members and Gun Violence

Gang communities and their members have been using Twitter to subdue their rivals, and identification of such users on Twitter facilitates the law enforcement agencies to anticipate the crime before it can happen. Balasuriya et al. [49] investigated conversations for finding street gang members on Twitter. A review of the profiles of gang members segregates them from rest of the Twitter population by checking hashtags, YouTube links, and emojis in their content [96]. In [49], nearly 400 gang member profiles were manually identified using seed terms, including gang affiliated rappers, their retweeters, followers as well as followees. They used textual features of the tweet, YouTube video descriptions and comments, emojis and profile pictures to power various machine learning algorithms including Naive Bayes, Logistic Regression, Random Forest and Support Vector Machines, to train the model. Random Forest performed well for Gang and Non-Gang classification. It is interesting to notice that gang members usually make use of their profile images in a specific way to intimidate other people and members of rival gangs.

As gun control policies in big cities, such as Chicago, have changed over the years, the volume of the taunting and threatening conversations on social media has also relatively increased [97]. Such conversations can be leveraged to assist law enforcement officers by providing insights on situational awareness as well as predicting a conflict between gang groups for a possible gun violence incident. Blevins et al. [97] used a Twitter dataset that was manually labeled by a team of researchers with expertise in cyber-bullying, urban-based youth violence and qualitative studies. Their strategy was to collect all tweets, mentions, replies, and retweets from a specific user profile between 29 March and 17 April 2014. Three experts developed the key types of content and used the work by Bushman and Huesmann [98] to identify and categorize types of aggression. To overcome the challenge of recognizing special slang terms and local jargons in tweets as mentioned in Desmond et al. work, Blevins et al.[97] developed a part-of-speech (POS) tagger for the gang jargon and mapped the vocabulary they use to Standard English using machine translation alignment. They developed emotion classifier that uses the extracted POS tags, and Dictionary of Affect in Language (DAL) quantitative scores (Whissell, 2009) as key features. Ternary classification is applied to the whole dataset (TCF) and binary classification on the aggression-loss subset (BCS). Then they use a cascading classifier (CC), which uses two SVM models. Initially, one SVM model is used to filter the tweets into aggression/loss tweets, and all other tweets fall into the other category. After this filtration, only aggression/loss tweets is passed to second SVM model which is again a binary classifier for loss or aggression. So this Aggression Supervised classifier is able to categorize loss with 62.3% F-score and aggression with 63.6% F-score which beats the baseline model (Unigrams) by 13.7 points (aggression) and 5.8 points (loss) [97].

4.5 Transportation

Congestion due to traffic is one of the prevalent problems in the United States (U.S.). Even after having structured rules that govern the flow of the traffic in the U.S., congestion due to non-recurring activities still affects the schedules of people. However, the stationing of police officers to smooth the traffic is a probable solution, although it would not be long-term. Having an estimation of the flow of traffic in the advent of an event can help people to re-route their path to the destination. Leveraging social media and machine learning to estimate traffic is one such long-term solution that can be drafted for active traffic monitoring. Social media is flooded with posts from people about an event. Such posts can provide the location of the event or the tweeter, and it can be used along with other textual features to estimate the traffic flow. In [99], textual features, tweet and user-metadata such as text, hashtags, URLs, number of users and retweeted tweets were used by combining with live event data to predict traffic dynamics. They utilized autoregressive model, neural network, support vector regression, and K-nearest neighbor for traffic prediction. The evaluation was performed using mean absolute percentage error (MAPE), and root means square error (RMSE), with support vector regression (SVR) performing better over other regression models. SVR reduced the error in traffic prediction by 24% in terms of RMSE.

4.6 Location Estimation

Social media serves a vital role in times when people struggle to survive a disastrous event such as hurricane or earthquake, to provide solutions for assisting the public in recovery efforts. These solutions include identification of the demand and its location, and mapping the identified demands with suitable suppliers analyzing Twitter data.

In particular, location extraction plays a significant role in identifying the area that is impacted by a disaster as well as providing assistance [100]. Mahmud et al. [101] developed an approach to predict the location of users at the city level on Twitter combining several classifiers. They removed stop words, performed part-of-speech tagging, extracted hashtags, and extracted a feature called local term, a term used by local people to refer to the city. For detecting the local terms, several classification algorithms and found Nave Bayes, SVM and Decision trees (J48) as the best performing algorithms. Al-Olimat et al. [102] developed a tool called LNEx (Location Name Extraction), that extracts the location from the tweet content by utilizing the OpenStreetMap[103], GeoNames [104] and DBpedia [105] for disambiguation. The information retrieval process from the tweet is two-fold, which are toponym extraction and geoparsing. Toponym is a process to extract city and street names, points of interest, from unstructured text, tweets in particular for this study. Location names are usually abbreviated on Twitter; hence, a text normalization procedure is used for expansion of such brevity. For instance, tweets may contain Rd as an abbreviation, and it is normalized to road. Furthermore, ambiguous location problems are resolved by employing the geoparsing procedure using the OpenStreetMap API[28]. LNEx improved the average F-Score by 98-145%, outperforming all the state of art taggers.

4.7 Community on Social Media

People with distinct feelings, expression, solutions, and intelligence, share their opinions on Twitter. Such a diversified content can be related to elections, football game or a domain that is influenced by public views. With the abundance of textual data, one can envision the power of collective intelligence that can be harnessed for a wise recommendation, judgment and strategy building. Also, it is a known fact that a judgment call made by a crowd is superior to an individuals decision [106]. Formation of a diverse group can improve the decision-making process through what is known as Wisdom of Crowd (WoC). WoC is meant to minimize regional biases that may cloud objectivity associated with individuals judgment and bring together different perspectives and knowledge that can enhance coverage and comprehensiveness of the analysis. For example, WoC can be used to design a portfolio of stocks that maximize the profit in the stock market trading. However, no existing work illustrates the notion of WoC statistically and analytically. A methodological way for measuring the diversity of the crowd is crucial to the rise of human social engagement on social media. According to a recent survey from Pew Research Center, 76% of the American population is active on social media. It attributes success to a significant amount of online data and can aid in creating WoC of the social system. In [107], fantasy premier league (FPL) is considered to exercise the better judgment of the diverse crowd. In their work, they predict the best performing team captain in the premier league, an element dictating the success of a team, based on the scores retrieved from the fantasy football and content of Twitter users. They utilized Word2vec similarity measure to quantify the diversity of two groups of users during captain selection in FPL. Furthermore, They defined and validated their statistical objective scoring criteria to measure the quality of crowd judgment.

4.8 Demographics

In many applications, demographic information is a key to analysis that depends on different segments of the population concerning age groups, ethnicity, and

[28] https://wiki.openstreetmap.org/wiki/API_v0.6

gender. For example, age is critical for understanding drug abuse, while gender is critical to understand vulnerability to depression. Twitter in its current state does not require users to provide any demographic information.

4.8.1 Age Estimation using social media

Researchers developed a machine learning system coupled with the DBpedia knowledge graph utilizing the user follower-followee networks to predict the most probable age of a Twitter user, in [108]. They gathered pre-identified famous people from DBpedia, based on their occupations and areas of interest, which also included their birth dates. Then they extracted a sample of 23,120 users who are in one/two hops of follower-followee network of famous people. Some of the user profiles were spam/bot and hence they were removed. Then they selected 16K users among the followers of the top 50 famous figures as their training set and 8K as their testing set. They achieved 84% accuracy in predicting the age of these users. They selected Support Vector Regression (SVR) with K-Fold Cross Validation [109] as their best performing model after evaluating using Linear Regression [110], Least Absolute Shrinkage and Selection Operator (LASSO) [111], and ElasticNet [112].

Zhang et al. [113] studied the problem of age prediction on Twitter, using SVM and least square optimization algorithm in building the model. They utilized various features such as linguistic, textual, and network, to improve their model, achieving an F1 score of 0.81. They discovered that the characteristics of users in the same age groups have similar content and interactions between each other. On the other hand, Nguyen et al. [114] investigated the relationship between the language used in tweets of a user and his/her age. They annotated the dataset that was collected following a guideline formed based on the tweet content of users in different age groups such as explicit or implicit age or life stage mentions. They found that the language use of people in same age groups is similar regarding the word and phrase selections as well as the topics that they are talking about. For instance, the following two sets of words, school, son, daughter, wish, enjoy, thanks, take care and haha, xd, internship, school have been used by users in two different age groups. In their analysis, they used linear and logistic regression models with unigram feature only, achieving an F1 score of 0.76.

4.8.2 Gender Estimation using Twitter

Estimation of the gender of a twitter user is beneficial to the analysis of Twitter data for health-related, drug abuse, and harassment activities. Existing approaches utilized statistical features [115] and seldom involved background knowledge along with social information. In [116], a dataset from Sina Weibo, which is a counterpart of the micro-blogging platform Twitter, in China, was used to assess their methodology for gender prediction. [117] exploits online behavioral and textual features and choice of vocabulary for each user. Online behavioral features include the number of fans, attention, messages, comments, forwards and a ratio of original/forwarded messages. Textual features include hashtags, URLs, emoticons, and sentence length. They also made use of username and pictures in content. Lexical features were extracted from the content using TF-IDF. They used four algorithms for predicting gender: Decision Tree, Naive Bayes, Logistic Regression and Support Vector Machines (SVMs), and found that SVM outperformed other classifiers by attaining accuracy of 94.3%.

4.9 Anomaly & Popularity Prediction

Twitter has become a playground for spammers. While public conversations on Twitter are diverse and challenging to analyze and summarize, spammers and bots further complicate the reliability of the outcome. Bots are automated software that is programmed to post a predefined content. They are being used mostly to propagate or promote bias and skew votes in politics, views on social issues, or provide impetus to promotional campaigns. On the other hand, prediction of the popularity of trending topics or issues requires robust analysis that takes into account anomalous accounts.

Thomas et al. [25] collected 1.8 billion tweets sent by 32.9 million users and manually identified 1.1 million suspended accounts as spammer accounts along with 80 million anomalous tweets. They used user behavior regarding interactions with other users, public Twitter handler service usage and textual features of tweet content such as shortened URLs created using free web hosting services. Volkova et al. [118] also studied this problem by applying a deep learning technique, Recurrent Neural Networks (RNN), using tweet metadata and network features. They compared their approach with state-of-the-art machine learning methods such as log-linear models. Their RNN model outperformed all the machine learning models built using various combinations of features with 0.95 F1 score. Sentiment has also been used in spam detection works [119, 120] as a feature to detect bots on Twitter. Varol et al. [120] also studied the detection of online bots on Twitter, and utilized Random Forests, AdaBoost, Logistic Regression and Decision Tree algorithms. They found Random Forest classifier achieved the best performance with 0.95 AUC score. They made use of sentiment features that they extracted from the text beside tweet and user metadata, textual, linguistic and network features.

Poblete et al. [121] have investigated the tweet credibility issue in the news disseminated on the platform. They crowdsourced the task of evaluating the credibility of each tweet to determine if it has newsworthy topics, labeling each tweet using automated credibility analysis. Labels given by crowdsourcing process were used in the training phase. They used SVM, decision trees, decision rules and Bayesian networks, and best results were given by J48 decision tree, achieving an 86% F1 score. Ross et al. [122] created a robust and general feature set for learning to rank tweets based on credibility and newsworthiness. In previous works by Gupta et al. [123–125], they have demonstrated that when the training and testing data are from two distinct time periods, the ranker performs poorly. Ross et al. [122] improved upon this by creating a feature set that does not overfit a particular year or a set of topics, which is critical for robust analysis of social media over time and across different domains.

Varol et al. [34] conducted a time series analysis to predict if a trending meme is organic or promoted by a group. They aimed to predict memes that have potential to trend before it becomes trending; therefore, the task of predicting trends is naturally forced to utilize a sparse dataset. For this reason, they had to reliably extract textual, linguistic, tweet and user metadata, network and statistical features, from a small dataset. They used three learning algorithms namely, K-Nearest Neighbor (KNN) with Dynamic Time Warping (KNN-DTW), Symbolic Aggregate approXimation with Vector Space Model (SAX-VSM) and KNN. KNN is a machine learning algorithm for classification and DTW for multi-dimensional time series. They found KNN-DTW and KNN showed the best performance in prediction. They used AUC as evaluation metric to measure accuracy because it is not biased by the imbalance in classes (e.g., 75 promoted trends versus 852 organic ones). Weng et al. [126] studied the prediction of the popularity of meme on Twitter. They relied mostly on network features besides tweet and user metadata, using random forest and linear regression. They extracted 13 features such as some early adopters, average shortest network path length between users, the diameter between users, and the number of infected communities. They built their model using random forest and tested against five different baselines that used linear regression along with different combinations of the 13 features. Their model achieved 0.85 F1 score, outperforming the baselines. Kobayashi et al. [127] predicted the popularity of a tweet in terms of the number of retweets in a time window in the future. They used time series analysis using a method called time-dependent Hawkes process (TiDeH) calculating infectious rate and using tweet and user metadata such as temporal information from a tweet and number of followers of a user. They evaluated their system against other existing methods that incorporated linear regression and Poisson process and reported that it outperformed other approaches achieving around 5% mean error rate. Tsur et al. [128] also studied the popularity of hashtags on Twitter, through linguistic features of the tweet text, specifically hashtags. They obtained promising results using a modified version of Gradient Boosted Trees called Gradient Boosted rank. They compared their approach with SVM and Least-effort algorithms, obtaining 0.11 mean error rate. Ruan et al. [129] predicted the volume of tweets, analyzing the user behavior on individual as well as collective level. Besides tweeting activity and content analysis of users, they utilized the underlying follower-followee network, user network structure, neighboring friends influence and user past activity as features. They used linear regression model with multiple features that include network structure, user interaction, content characteristics and past activity, and found that combining features yields the best performance.

4.10 Sales & Stock Price Prediction

As social media, particularly Twitter, users share their satisfaction or frustration with products on the platform, these user reviews can be exploited by companies to generate actionable insights to meet customer expectations and eventually provide better quality products and services. Industrial applications of predictive analysis of social media have been gradually adopted, to gain the understanding

Table 1. Comparative Analysis of Applications and their Evaluation. Acronyms for Algorithms and Features are described in Table 2

Ref. & Evaluation	Application	Algorithms	Features
[39] F1=0.88 [113] F1=0.81 [130] F1=0.59 [131] F1=0.83 [49] F1=0.77 [132] Acc=0.82 [89] Acc=0.92 [90] F1=∼0.75 [114] F1=0.76 [54] AUC=∼0.7 [15] AUC=0.8 [108] Acc=0.84	User Profiling User Classification	SVM LinR CNN RF NB LogR, LASSO	UsM, TwM Ling, Nw, Stat, Txt Vis
[53] Acc=0.72 [133] F1=0.62	User Attitude, Personality, and Mood Prediction	SVM, NB, RF	TwM, Txt, Ling, Nw, Stat
[134] AUC=0.8	Sales & Stock price prediction	NB, RF, SVM	TwM
[59] Med error=0.32 [135]] F1=0.58 [136] Acc=0.85 [137] AUC=0.91	Social and Political events, Elections, Collective action	PosR NBR, SVM, LogR, CNN, RF	TwM, UsM Txt, Ling, Nw
[34] AUC=0.95 [126] F1=∼0.85 [127] Mean err= ∼0.179 [128] Mean err=∼0.11	Popularity prediction	KNN-DTW, SAX-VSM, KNN, RF, TiDeH, LinR, LogR GrB, SVM	Txt Nw and Stat, TwM and UsM Ling
[25] Acc=0.89 [34] AUC=0.95 [119] AUC=0.73 [118] F1= 0.95 [138] F1=0.99 [120] AUC=0.95 [121] F1=0.86	Spam bot detection Troll detection Credibility prediction	KNN-DTW, SAX-VSM, KNN, LinR, DT NB, GrB RF, AdB, BNet, RNN, SVM, LogR	Text, Nw, Ling, TwM, UsM Stat, Vis
[17] Pre=∼0.80 [16] F1=0.77 [18] F1=0.67 [22] F1=∼0.60 [29] F0.5=0.62 [21] Acc=0.82 [6] Acc=0.61 [139] Mean error=0.071 [140] Acc=∼0.71 [141] F0.5= 0.76 [64] F1=∼0.73 [26] F1=0.79	Sentiment analysis Emotion Detection	NB, SVM, CRF, LibLin, RBF-NN, AdB, RT, REPTree, BNet, LogR	Ling, Txt, TwM and UsM
[101] Acc=∼0.83 [142] Mean error=0.39 [143] Acc=0.88 [144] Acc=0.79	Location Estimation Traffic Estimation	NB, SVM, DT, LinR, MxEnt	Ling, TwM, UsM Txt
[145] F1= 0.70 [146] Pre=0.86	Finding Important Users, Community Detection	RNN, SVM	Nw, UsM, Txt
[147] LFK-NMI=0.13 [148] F1=0.63	Topic Extraction, Meme Extraction	HiCl, KM, LDA, SVM	Txt, Ling, TwM, UsM Nw
[149] Acc=0.84 [150] AUC=0.83 [151] F= 0.8736 [83] K Alpha=0.84	Public Health Health-care	NB, SVM, RF LDA, ssToT	Txt, Ling, TwM Sentiment,
[102] F1=0.81 [152] R2= 0.67	Disaster Management	LogR	Txt, Ling, UsM, TwM Nw

of the market. Some of the use-cases that have adopted social media data for decision making are for:

1. Improvement of Customer Service: Delta Airline exploited social media to identify the reasons for customer frustration. For instance, lost luggage or poor service.
2. New Products Research and Development: JD Power quality assessment has determined that car company modify car seats based user sentiments on the social sphere [153].
3. Key Influencers: A cosmetics company L'Oreal uses social media follower-followee network to find Influencers for promotions[29].
4. Recommendations through deep learning: YouTube utilizes the deep neural network to enhance their recommendation system using implicit feedback by analyzing users comments and videos of interest [154].

Georgiev et al. [134] investigated the question of how the Olympic Games impacted the sales of businesses in London. They used Twitter posts along with the check-ins through Foursquare platform to extract mostly location-specific features from tweet text and tweet metadata, such as the distance of businesses to stadiums and sponsor businesses, transitions to entertainment places and social areas. They evaluated their work using AUC, for Nave Bayes, Random Forest, and SVM algorithms and reported that SVM performed best with 0.8 of AUC.

[155] employs feed-forward network (FF) for predicting the likelihood of a customer to buy a product. They restricted their dataset to tweets about mobile phones and cameras, expensive products that people often buy after doing some research online. Before predicting the likelihood of purchasing a product, they predicted whether a tweet represents the respective users purchasing behavior. Then they predict whether the user will purchase the product after 60 days time window of tweeting. They compared the performance of their approach with Long Short Term Memory (LSTM), Recurrent Neural Networks (RNNs) (with varying dropout rates) based implementation and observed that their approach with FF surpasses others by small margins. FF learning cycle involved RM-Sprop [156], sigmoid activations and negative log-likelihood function with batch training.

5 Conclusion

Twitter has positioned itself as an essential part of the social media environment becoming an emergent communication medium. This development has opened up new opportunities for researchers to gauge the pulse of the populace reliably and use that to study public opinion, form policies, understand the impact of events, and find newer ways to address certain problems. Social media data has already enabled researchers to predict the trends and outcomes of several critical real-world events, and its reliability and coverage can further be improved

[29] https://www.socialmediatoday.com/special-columns/adhutchinson/2015-09-09/big-brand-theory-loreal-stays-connected-their-audience

by incorporating background knowledge [85, 20]. Specifically, monitoring the engagement and public opinion about ongoing events from temporal and spatial perspectives can foretell their evolution as well as the outcome. Moreover, this information can complement traditional surveys or polls that are conducted by non-government agencies to improve our confidence in the prediction, as traditional methods alone can be misleading or sluggish in reacting to rapidly changing events. In order to account for the complex decision making that requires consideration of a number of factors that can impact a situation or an event, incorporation of as many signals as possible in comprehending the big picture is necessary. We have explored a predictive analysis paradigm that comprises two levels of prediction, using coarse-grained analysis built upon fine-grained analysis. Such analysis have been conducted with creditable success for events such as elections, gun violence, drug misuse or illicit drug use [3].

In this chapter, we have discussed processes, algorithms, and applications concerning predictive analysis in different domains. We illustrated fine-grained analysis by customizing domain-independent approaches to extract signals such as sentiment, emotions, and topics through the application of machine learning models, and coarse-grained analysis by aggregating and cultivating the signals to make predictions. We have also provided details of related prominent studies in ten different domains such as healthcare, public health, political and social issues, disaster management, sales and stock prediction, and demographics. The following table summarizes related work describing various applications and methods used.

Table 2. The Acronyms used in the comparative table.

Acronym	**Algo. Description**	**Acronym**	**Feature Descr.**
LinR	Linear Regression	UsM	User metadata
RF	Random Forest	TwM	Tweet metadata
NB	Nave Bayes	Ling	linguistic
LogR	Logistic Regression	Nw	Network
PosR	Poisson Regression	Stat	Statistical
NBR	Negative Binomial Reg.	Txt	Textual
GB	Gradient Boosting	Vis	Visual
AdB	AdaBoost		
DT	Decision Trees		
BNet	Bayes Net		
LibLin	LIBLINEAR		
HiCl	Hierarchical Clustering		
KM	K-Means		
RT	Random Tree		

6 Acknowledgement

We are grateful to Amelie Gyrard, Mustafa Nural, Sanjaya Wijeratne, Shreyansh Bhatt, Ankita Saxena for their assistance with their reviews and comments that greatly improved this book chapter.

We acknowledge partial support from the National Science Foundation (NSF) award CNS-1513721: "Context-Aware Harassment Detection on Social Media", National Institutes of Health (NIH) award: MH105384-01A1: ”Modeling Social Behavior for Healthcare Utilization in Depression”, NSF award EAR- 1520870 Hazards SEES: Social and Physical Sensing Enabled Decision Support for Disaster Management and Response, Community in Social Media : This work was supported by Army Research Office Grant No. W911NF-16-1-0300, National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02 Trending: Social media analysis to monitor cannabis and synthetic cannabinoid use. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, NIDA or Army Research Office.

References

1. H. Purohit and A. Sheth, “Twitris v3: From Citizen Sensing to Analysis, Coordination and Action,” in *ICWSM*, 2013.
2. C. A. Davis, G. L. Ciampaglia, L. M. Aiello, K. Chung, M. D. Conover, E. Ferrara, A. Flammini, G. C. Fox, X. Gao, B. Gonçalves, P. A. Grabowicz, K. Hong, P.-M. Hui, S. Mccaulay, K. Mckelvey, M. R. Meiss, S. Patil, C. P. Kankanamalage, V. Pentchev, J. Qiu, J. Ratkiewicz, A. Rudnick, B. Serrette, P. Shiralkar, O. Varol, L. Weng, T.-L. Wu, A. J. Younge, and F. Menczer, “OSoMe: the IUNI observatory on social media,” *PeerJ Computer Science.*
3. A. Sheth, H. Purohit, G. A. Smith, J. Brunn, A. Jadhav, P. Kapanipathi, C. Lu, and W. Wang, “Twitris: A System for Collective Social Intelligence,” *Encyclopedia of Social Network Analysis and Mining*, 2018.
4. K. B. Penuel and M. Statler, *Encyclopedia of disaster relief.* Sage Publications, 2011.
5. J. Malilay, M. Heumann, D. Perrotta, A. F. Wolkin, A. H. Schnall, M. N. Podgornik, M. A. Cruz, J. A. Horney, D. Zane, R. Roisman, J. R. Greenspan, D. Thoroughman, H. A. Anderson, E. V. Wells, and E. F. Simms, “The Role of Applied Epidemiology Methods in the Disaster Management Cycle,” *American Journal of Public Health*, vol. 104, no. 10, pp. 2092–2102, 2014.
6. W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, “Harnessing Twitter ’Big Data’ for Automatic Emotion Identification,” in *IEEE International Confernece on Social Computing (SocialCom)*, 2012.
7. F. R. Lamy, R. Daniulaityte, R. W. Nahhas, M. J. Barratt, A. G. Smith, A. Sheth, S. S. Martins, E. W. Boyer, and R. G. Carlson, “Increases in synthetic cannabinoids-related harms: Results from a longitudinal web-based content analysis,” *International Journal of Drug Policy*, 2017.
8. A. Sheth and P. Kapanipathi, “Semantic Filtering for Social Data,” *IEEE Internet Computing*, 2016.
9. P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant, “Personalized Filtering of the Twitter Stream,” in *SPIM Workshop at ISWC 2011*, 2011.
10. P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, “User Interests Identification on Twitter Using a Hierarchical Knowledge Base,” in *European Semantic Web Conference*, 2014.
11. D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins, and R. Falck, “PREDOSE: A semantic web platform for drug abuse epidemiology using social media,” *Journal of Biomedical Informatics*, vol. 46, pp. 985–997, 2013.
12. H. Saif, *Semantic Sentiment Analysis in Social Streams*, 2017.

13. S. Wijeratne, A. Sheth, S. Bhatt, L. Balasuriya, H. S. Al-Olimat, M. Gaur, A. H. Yazdavar, and K. Thirunarayan, "Feature Engineering for Twitter-based Applications," in *Feature Engineering for Machine Learning and Data Analytics*, 2017, p. 35.
14. K. Gimpel, N. Schneider, B. O 'connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," *Proceedings of ACL*, 2011.
15. C. Wagner, S. Asur, and J. Hailpern, "Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter," in *SocialCom*, 2013.
16. X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach," in *Proceedings of the 20th ACM international conference on Information and knowledge management. ACM*, 2011.
17. D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," in *Proceedings of the 23rd international conference on computational linguistics. ACM*, 2010, pp. 241–249.
18. E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, pp. 538–541, 2011.
19. D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter," in *Proceedings of the 20th international conference on World wide web.*, 2011.
20. F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," in *ICWSM*, 2013, pp. 400–408.
21. A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Tech. Rep., 2009.
22. A. Agarwal, B. Xie, and I. Vovsha, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, no. June, 2011, pp. 30–38.
23. B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network," in *IEEE international conference on Social Computing Social computing (SocialCom)*, 2010.
24. N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad News Travel Fast : A Content-based Analysis of Interestingness on Twitter," in *Proceedings of the 3rd International Web Science Conference. ACM*, 2011.
25. K. Thomas, C. Grier, and V. Paxson, "Suspended Accounts in Retrospect: An Analysis of Twitter Spam," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement*, 2011.
26. K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon Smoothed Language Models for Twitter Sentiment Analysis," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
27. C. Zhai, J. Lafferty, J. Lafferty, and C. Zhai, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 179–214, 2004.
28. M. Boia and B. Faltings, "A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets," in *SocialCom*, 2013.
29. A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *LREc*, vol. 10, 2010.

30. R. Kelly and L. Watts, “Characterising the Inventive Appropriation of Emoji as Relationally Meaningful in Mediated Close Personal Relationships,” *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*, 2015.
31. P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of Emojis,” *PLOS One*, 2015.
32. H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, “Blissfully happy or ready to fight: Varying Interpretations of Emoji,” *International AAAI Conference on Web and Social Media*, no. ICWSM, pp. 259–268, 2016.
33. S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran, “EmojiNet: An Open Service and API for Emoji Sense Discovery,” in *ICWSM*, 2017.
34. O. Varol, E. Ferrara, F. Menczer, and A. Flammini, “Early detection of promoted campaigns on social media,” *EPJ Data Science*, vol. 6, no. 1, 2017.
35. L. C. Freeman, “Centrality in Social Networks Conceptual Clarification,” *Social Networks*, vol. 179, pp. 215–239, 1978.
36. L. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
37. P. Bonacich, “Power and centrality : A family of measures,” *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.
38. G. Lawyer, “Understanding the influence of all nodes in a network,” *Nature Scientific Reports*, 2015.
39. M. Pennacchiotti and A.-M. Popescu, “Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.
40. R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, N. Tziritas, C.-Z. Xu, A. Y. Zomaya, A. S. Alzahrani, and H. L, “A Survey on Text Mining in Social Networks,” *The Knowledge Engineering Review*, vol. 000, pp. 1–24, 2004.
41. A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, D. Chek, and L. Ngo, “Text mining for market prediction: A systematic review,” *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 41, pp. 7653–7670, 2014.
42. F. Franch, “(Wisdom of the Crowds) : 2010 UK Election Prediction with Social Media,” *Journal of Information Technology & Politics*, vol. 10, no. 1, pp. 57–71, jan 2013.
43. F. Bravo-Marquez, D. Gayo-Avello, M. Mendoza, and B. Poblete, “Opinion Dynamics of Elections in Twitter,” in *Eighth Latin American Web Congress*, 2012.
44. L. Hong, O. Dan, and B. Davison, “Predicting Popular Messages in Twitter,” in *WWW*, 2011.
45. M. Sokolova, K. Huang, S. Matwin, J. Ramisch, V. Sazonova, R. Black, C. Orwa, S. Ochieng, and N. Sambuli, “Topic Modelling and Event Identification from Twitter Textual Data,” *ArXiv preprint*, 2016.
46. S. T. Dumais, “Latent semantic analysis,” *Annual Review of Information Science and Technology*, vol. 3, no. 11, p. 4356, 2008.
47. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Advances in neural information processing systems*, 2013.
48. S. Wijeratne, L. Balasuriya, D. Doran, A. Sheth, and A. Org, “Word Embeddings to Enhance Twitter Gang Member Profile Identification,” in *IJCAI Workshop on Semantic Machine Learning*, 2016.

49. L. Balasuriya, S. Wijeratne, D. Doran, and A. Sheth, "Finding Street Gang Members on Twitter," in *ASONAM*, 2016.
50. S. Sakaki, Y. Miura, X. Ma, K. Hattori, and T. Ohkuma, "Twitter User Gender Inference Using Combined Analysis of Text and Image Processing," in *Proceedings of the 25th International Conference on Computational Linguistics*, 2014, pp. 54–61.
51. K. Bontcheva, L. Derczynski, A. Funk, M. a. Greenwood, D. Maynard, and N. Aswani, "TwitIE : An Open-Source Information Extraction Pipeline for Microblog Text," *Proceedings of Recent Advances in Natural Language Processing*, no. September, pp. 83–90, 2013.
52. T. Mitra and E. Gilbert, "CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations," in *ICWSM*, 2016.
53. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," in *ICWSM*, 2013.
54. Y. Lewenberg, Y. Bachrach, and S. Volkova, "Using emotions to predict user interest areas in online social networks," in *Data Science and Advanced Analytics (DSAA)*, 2015.
55. H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2012, pp. 115–120.
56. M. Ebrahimi, A. H. Yazdavar, and A. Sheth, "On the Challenges of Sentiment Analysis for Dynamic Events," *IEEE Intelligent Systems*, 2017.
57. S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness," in *CHI - Crisis Informatics*, 2010.
58. L. Chen, W. Wang, and A. P. Sheth, "Are Twitter Users Equal in Predicting Elections? A Study of User Groups in Predicting 2012 U.S. Republican Presidential Primaries," in *Social Informatics*, 2012.
59. M. De Choudhury, S. Jhaver, B. Sugar, and I. Weber, "Social Media Participation in an Activist Movement for Racial Equality," in *ICSWM*, no. Icwsm, 2016, pp. 92–101.
60. H. Purohit, A. Hampton, V. L. Shalin, A. P. Sheth, J. Flach, and S. Bhatt, "What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination," *Computers in Human Behavior*, vol. 29, pp. 2438–2447, 2013.
61. H. Purohit, A. Hampton, S. Bhatt, V. L. Shalin, A. P. Sheth, and J. M. Flach, "Identifying Seekers and Suppliers in Social Media Communities to Support Crisis Coordination," in *Computer Supported Cooperative Work (CSCW)*, 2014.
62. H. Purohit, S. Bhatt, A. Hampton, V. L. Shalin, and A. P. Sheth, "With Whom to Coordinate, Why and How in Ad- Hoc Social Media Communications during Crisis Response," in *Proceedings of the 11 th International ISCRAM Conference*, 2014, pp. 787–791.
63. Shreyansh Bhatt, Hemant Purohit, and Andrew Hampton, "Assisting Coordination during Crisis: A Domain Ontology based Approach to Infer Resource Needs from Tweets," in *Web Science*, 2014.
64. L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, 2012, pp. 6:1–6:8.

65. D. Stojanovski, G. Strezoski, G. Madjarov, and I. Dimitrovski, “Finki at SemEval-2016 Task 4: Deep Learning Architecture for Twitter Sentiment Analysis,” in *Proceedings of SemEval*, 2016, pp. 149–154.
66. A. Esuli, F. Sebastiani, C. Nazionale, and D. Ricerche, “Optimizing Text Quantifiers for Multivariate Loss Functions,” *ACM Transactions on Knowledge Discovery from Data ACM Trans. Knowl. Discov. Data. VV*, vol. 26, 2015.
67. T. L. Griffiths, M. Steyvers, J. B. Tenenbaum, and T. Griffiths, “Topics in semantic representation Topics in semantic representation,” *Psychological review*, 2007.
68. L. Chen, C. Org, W. Wang, W. Org, M. Nagarajan, S. Wang, A. P. Sheth, and A. Org, “Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter,” in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
69. L. Hong and B. D. Davison, “Empirical Study of Topic Modeling in Twitter,” in *1st Workshop on Social Media Analytics (SOMA ’10)*, 2010.
70. W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, “Topical Keyphrase Extraction from Twitter,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 379–388.
71. X. Wang, M. S. Gerber, and D. E. Brown, “Automatic Crime Prediction using Events Extracted from Twitter Posts,” in *International conference on social computing, behavioral-cultural modeling, and prediction. Springer*, 2012.
72. N. Bhattacharya, I. Arpinar, and U. Kursuncu, “Real Time Evaluation of Quality of Search Terms during Query Expansion for Streaming Text Data Using Velocity and Relevance,” in *Proceedings - IEEE 11th International Conference on Semantic Computing, ICSC 2017*, 2017.
73. L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, “Using Social Media To Predict the Future: A Systematic Literature Review,” in *arXiv preprint*, 2017.
74. J. M. Robillard, T. W. Johnson, C. Hennessey, B. L. Beattie, and J. Illes, “Aging 2.0: Health Information about Dementia on Twitter,” *Plos One*, vol. 20, no. 87, 2013.
75. V. M. Prieto, S. Rgio Matos, M. Lvarez, F. Cacheda, J. L. Oliveira, and J. A. Añ, “Twitter: A Good Place to Detect Health Conditions,” *PLoS ONE*, vol. 9, no. 1, 2014.
76. A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, “Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media,” in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2017.
77. G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead Ihmc, “From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses,” in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 1–10.
78. T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis, “Multiview Deep Learning for Predicting Twitter Users’ Location,” *arXiv preprint*, 2017.
79. J. H. Lau and T. Baldwin, “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation,” *arXiv preprint*, 2016.
80. A. Grover and J. Leskovec, “node2vec: Scalable Feature Learning for Networks,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.

81. H. Bo, P. Cook, T. Imoth, and B. Dw, “Geolocation Prediction in Social Media Data by Finding Location Indicative Words,” in *Proceedings of COLING 2012*, 2012, pp. 1045–1062.
82. R. Daniulaityte, R. W. Nahhas, S. Wijeratne, R. G. Carlson, F. R. Lamy, S. S. Martins, E. W. Boyer, G. A. Smith, and A. Sheth, “”Time for dabs”: Analyzing Twitter data on marijuana concentrates across the U.S. HHS Public Access,” *Drug Alcohol Depend*, vol. 155, pp. 307–311, 2015.
83. F. R. Lamy, R. Daniulaityte, A. Sheth, R. W. Nahhas, S. S. Martins, E. W. Boyer, and R. G. Carlson Francois R Lamy, “”Those edibles hit hard”: Exploration of Twitter data on cannabis edibles in the U.S HHS Public Access,” *Drug Alcohol Depend*, vol. 1, no. 164, pp. 64–70, 2016.
84. P. N. Howard, M. Hussain, and W. Mari, “Opening Closed Regimes What Was the Role of Social Media During the Arab Spring?” 2011.
85. Z. Tufekci, “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls,” in *ICWSM*, 2014.
86. I. Arpinar, U. Kursuncu, and D. Achilov, “Social media analytics to identify and counter islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online,” in *Proceedings - 2016 International Conference on Collaboration Technologies and Systems, CTS 2016*, 2016.
87. G. Haciyakupoglu and W. Zhang, “Social Media and Trust during the Gezi Protests in Turkey,” *Journal of Computer-Mediated Communication*, vol. 20, no. 4, pp. 450–466, 2015.
88. H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” *Journal of Economic PerspectivesVolume*, vol. 31, no. 2Spring, pp. 211–236, 2017.
89. T.-A. Hoang, W. W. Cohen, E.-P. Lim, D. Pierce, and D. P. Redlawsk, “Politics, Sharing and Emotion in Microblogs,” in *ASONAM*, 2013.
90. A. Makazhanov and D. Rafiei, “Predicting Political Preference of Twitter Users,” *Social Network Analysis and Mining*, 2014.
91. R. Cohen and D. Ruths, “Classifying Political Orientation on Twitter: It’s Not Easy!” in *ICWSM*, 2013.
92. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, “Learning from Bullying Traces in Social Media,” in *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 656–666.
93. Y. Chen, S. Zhu, Y. Zhou, and H. Xu, “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety,” in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012.
94. V. Edupuganti, “Harassment Detection on Twitter using Conversations,” Ph.D. dissertation, 2017.
95. R. Kandakatla, “Identifying Offensive Videos on YouTube,” Ph.D. dissertation, 2016.
96. S. Wijeratne, D. Doran, A. Sheth, and J. L. Dustin, “Analyzing the Social Media Footprint of Street Gangs,” in *Intelligence and Security Informatics (ISI)*, 2015.
97. T. Blevins, R. Kwiatkowski, J. Macbeth, K. Mckeown, D. Patton, and O. Rambow, “Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2196–2206.
98. B. Bushman and L. Huesmann, “Short-term and long-term effects of violent media on aggression in children and adults,” *Arch Pediatr Adolesc Med*, 2006.

99. M. Ni, Q. He, and J. Gao, "Using Social Media to Predict Traffic Flow under Special Event Conditions," in *The 93rd Annual Meeting of Transportation Research Board*, 2014.
100. R. Krishnamurthy, P. Kapanipathi, A. P. Sheth, K. Thirunarayan, and A. Sheth, "Location Prediction of Twitter Users using Wikipedia," 2014.
101. J. Mahmud, J. Nichols, and C. Drews, "Where Is This Tweet From? Inferring Home Locations of Twitter Users," in *ICWSM*, 2012.
102. H. S. Al-Olimat, K. Thirunarayan, V. Shalin, and A. Sheth, "Location Name Extraction from Targeted Text Streams using Gazeeer-based Statistical Language Models," *arxiv preprint*, vol. 11, no. 17, 2017.
103. M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, 2008.
104. D. Ahlers, "Assessment of the accuracy of geonames gazetteer data," in *GIR*, 2013.
105. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "Dbpedia a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 1, pp. 1–5, 2012.
106. M. D. Lee and M. N. Lee, "The relationship between crowd majority and accuracy for binary decisions," *Judgment and Decision Making*, vol. 12, no. 4, pp. 328–343, 2017.
107. S. Bhatt, B. Minnery, S. Nadella, B. Bullemer, V. Shalin, and A. Sheth, "Enhancing crowd wisdom using measures of diversity computed from social media data," in *Proceedings of the International Conference on Web Intelligence*, 2017.
108. A. Smith and M. Gaur, "What's my age?: Predicting Twitter User's Age using Influential Friend Network and DBpedia," *arxiv preprint*, 2018.
109. P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," *Encyclopedia of Database Systems. Springer*, 2009.
110. D. Nguyen, N. A. Smith, and C. P. Rosé, "Author Age Prediction from Text using Linear Regression," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics*, 2011.
111. C. Chen, Y. Chang, K. Ricanek, and Y. Wang, "Face age estimation using model selection," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 93–99, 2010.
112. A. Culotta, N. Kumar Ravi, and J. Cutler, "Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data," *Journal of Artificial Intelligence Research*, vol. 55, pp. 389–408, 2016.
113. J. Zhang, X. Hu, Y. Zhang, and H. Liu, "Your age is no secret: Inferring microbloggers' ages via content and interaction analysis," *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, no. Icwsm, pp. 476–485, 2016.
114. D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, ""How Old Do You Think I Am?": A Study of Language and Age in Twitter," in *ICWSM*, 2013.
115. D. Bamman, J. Eisenstein, and T. Schnoebelen, "GENDER IN TWITTER: STYLES, STANCES, AND SOCIAL NETWORKS," *CoRR*, 2012.
116. W. Li and M. Dickinson, "Gender Prediction for Chinese Social Media Data," in *Proceedings of Recent Advances in Natural Language Processing*, 2017, pp. 438–445.

117. L. Li, M. Sun, and Z. Liu, "Discriminating Gender on Chinese Microblog: A Study of Online Behaviour, Writing Style and Preferred Vocabulary," in *10th International Conference on Natural Computation (ICNC)*, 2014.

118. S. Volkova and E. Bell, "Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter across Languages," in *ICWSM, Association for the Advancement of Artificial Intelligence*, no. Icwsm, 2017, pp. 290–298.

119. J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, "Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?" in *ASONAM*, 2014.

120. O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," in *ICWSM*, 2017.

121. B. Poblete, C. Castillo, and M. Mendoza, "Information Credibility on Twitter," pp. 45–59, 2011.

122. J. Ross and K. Thirunarayan, "Features for Ranking Tweets Based on Credibility and Newsworthiness," in *International Conference on Collaboration Technologies and Systems*, 2016.

123. A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: A Real-timeWeb-based System for Assessing Credibility of Content on Twitter," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8851, no. November, 2014.

124. A. Gupta and P. Kumaraguru, "Credibility Ranking of Tweets during High Impact Events," in *PSOSM*, 2012.

125. A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy," in *WWW*, 2013.

126. L. Weng, F. Menczer, and Y.-Y. Ahn, "Predicting Successful Memes using Network and Community Structure," in *IC*, 2014, pp. 535–544.

127. R. Kobayashi and R. Lambiotte, "TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics," no. ICWSM, 2016, pp. 191–200.

128. O. Tsur and A. Rappoport, "Don't Let Me Be #Misunderstood: Linguistically Motivated Algorithm for Predicting the Popularity of Textual Memes," in *ICWSM, Ninth International AAAI Conference on Web and Social Media*, 2015, pp. 426–435.

129. Y. Ruan, H. Purohit, D. Fuhry, S. Parthasarathy, A. P. Sheth, and A. Sheth, "Prediction of Topic Volume on Twitter," in *4th International ACM Conference on Web Science*, 2012, pp. 397–402.

130. N. Pattisapu, M. Gupta, P. Kumaraguru, and V. Varma, "Medical Persona Classification in Social Media," in *ASONAM*, 2017.

131. Z. Gilani, E. Kochmar, and J. Crowcroft, "Classification of Twitter Accounts into Automated Agents and Human Users," in *ASONAM*, 2017.

132. J. S. Alowibdi, U. A. Buy, P. S. Yu, and L. Stenneth, "Detecting Deception in Online Social Networks," in *ASONAM*, 2014.

133. J. Mahmud, G. Fei, A. Xu, A. Pal, and M. Zhou, "Predicting Attitude and Actions of Twitter Users," in *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI '16*. ACM Press, 2016, pp. 1–6.

134. P. Georgiev, A. Noulas, and C. Mascolo, "Where Businesses Thrive: Predicting the Impact of the Olympic Games on Local Retailers through Location-based Services Data," in *ICWSM*, 2014, pp. 151–160.

135. X. Yang, R. Mccreadie, C. Macdonald, and I. Ounis, "Transfer Learning for Multi-language Twitter Election Classification," in *ASONAM*, 2017.

136. R. Korolov, D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, L. Kaplan, W. Wallace, J. Han, and H. Ji, "On Predicting Social Unrest Using Social Media," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.
137. N. Kallus, "Predicting Crowd Behavior with Big Public Data," *arXiv preprint*, 2014.
138. J. Echeverria and S. Zhou, "Discovery, Retrieval, and Analysis of the 'Star Wars' Botnet in Twitter," in *ASONAM*, 2017.
139. W. Gao and F. Sebastiani, "Tweet Sentiment: From Classification to Quantification," in *ASONAM*, 2015.
140. A. Hassan, A. Abbasi, and D. Zeng, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework," in *SocialCom*, 2013.
141. A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei, "Detecting Comments on News Articles in Microblogs," in *ICWSM*, 2013.
142. T. Georgiou, A. E. Abbadi, X. Yan, and J. George, "Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application," in *ASONAM*, 2015.
143. A. J. Aiswal, W. Peng, and T. Sun, "Predicting Time-sensitive User Locations from Social Media," in *ASONAM*, 2013.
144. D. Rout, D. Preoiuc-Pietro, K. Bontcheva, and T. Cohn, "Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties," in *24th ACM Conference on Hypertext and Social Media*, Paris, France, 2013.
145. B. Rath, W. Gao, J. Ma, and J. Srivastava, "From Retweet to Believability: Utilizing Trust to Identify Rumor Spreaders on Twitter," in *ASONAM*, 2017.
146. I. Bizid, N. Nayef, P. Boursier, S. Faiz, and J. Morcos, "Prominent Users Detection during Specific Events by Learning On-and Off-topic Features of User Activities," in *ASONAM*, 2015.
147. E. Ferrara, M. Jafariasbagh, O. Varol, V. Qazvinian, F. Menczer, and A. Flammini, "Clustering Memes in Social Media," in *ASONAM*, 2013.
148. S. Yamamoto and T. Satoh, "Hierarchical Estimation Framework of Multi-Label Classifying: A Case of Tweets Classifying into Real Life Aspects," in *ICWSM*, 2015.
149. A. Beykikhoshk, O. Arandjelovi, D. Phung, and S. Venkatesh, "Data-Mining Twitter and the Autism Spectrum Disorder: A Pilot Study," in *ASONAM*, 2014.
150. Z. Yin, Y. Chen, D. Fabbri, J. Sun, and B. Malin, "#PrayForDad: Learning the Semantics Behind Why Social Media Users Disclose Health Information," in *ICWSM*, 2016.
151. R. Daniulaityte, L. Chen, F. R. Lamy, R. G. Carlson, K. Thirunarayan, and A. Sheth, "When Bad' is Good': Identifying Personal Comm unication and Sentiment in Drug-Related Tweets," *JMIR PUBLIC HEAL TH AND SURVEILLANCE*, 2016.
152. Y. Hu, S. Farnham, and K. Talamadupula, "Predicting User Engagement on Twitter with Real-World Events," in *ICWSM*, 2015.
153. J. S. Kessler, M. Eckert, L. Clark, and N. J. Nicolov Power, "The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain," in *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*, 2010.
154. P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

155. M. Korpusik, S. Sakaki, F. Chen, and Y.-Y. Chen, “Recurrent neural networks for customer purchase prediction on twitter.” in *CBRecSys@ RecSys*, 2016, pp. 47–50.
156. T. Tieleman and G. Hinton, “Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. technical report.”