Feature Engineering for Twitter-based Applications

Chapter · October 2017

DOI: 10.1201/9781315181080-14

8 authors, including:

Sanjaya Wijeratne
Holler Technologies Inc.
27 PUBLICATIONS **544** CITATIONS
SEE PROFILE

Amit Sheth
University of South Carolina
1,018 PUBLICATIONS **42,772** CITATIONS
SEE PROFILE

Shreyansh Bhatt
Wright State University
25 PUBLICATIONS **418** CITATIONS
SEE PROFILE

Lakshika Balasuriya
Wright State University
13 PUBLICATIONS **403** CITATIONS
SEE PROFILE

Chapter 1

Feature Engineering for Twitter-based Applications

Sanjaya Wijeratne, Amit Sheth, Shreyansh Bhatt, Lakshika Balasuriya, Hussein S. Al-Olimat, Manas Gaur, Amir Hossein Yazdavar, Krishnaprasad Thirunarayan

Kno.e.sis Center, Wright State University, Dayton, OH, USA

1.1 Introduction

Social media websites have become extremely popular among online users in recent years. Surveys performed by Pew Research Center in 2016 claimed that social networking sites are visited by 69% of the total U.S. population

where 76% of them daily check those websites[1]. These online activities generate large amounts of user-generated content that can be mined to understand user interests and recommend products to online users, develop targeted marketing campaigns for products, understand the user's perspectives of a product, etc. Among many online social networking websites, Twitter has gained popularity due to the fact that users can follow any other user's activities, by accessing their short text messages, called 'tweets', posted to the Twitter network. For example, Twitter users can follow their favorite celebrities to learn what they share publically, in real-time. Currently, Twitter has grown to a social network of 328 million active users who post around 500 million messages collectively everyday[2].

With this rapid growth, Twitter has become a useful source for researchers and application developers to conduct studies and develop applications that analyze the pulse and nature of the populace. Researchers have studied the content shared on Twitter to understand the demographics of Twitter users [57, 42], their preference for a selected product or service [57], their sentiments [14], and emotions [79], etc. By letting Twitter users to share text, images, and video, Twitter provides an environment that supports sharing of multi-modal data. In order to glean insights from tweets, it is critical to design and learn suitable features from the raw data and rank order them based on their effectiveness for analysis or predictive task at hand. In the past, researchers have studied feature selection for specific problems such as sentiment analysis [45, 61], and rumor detection[75], where they have used Twitter as a data source. These studies focus only on a specific task at hand, but, these features can be reused or generalized for addressing diverse problems. Thus, identifying features that can work well on a range of Twitter-related applications could help researchers in feature selection and engineering tasks.

This chapter presents an analysis of feature engineering for Twitter-based applications. We begin with a discussion of how Twitter data can be downloaded from the Twitter Application Programming Interface (API). Then we describe different types of data available in tweets downloaded from Twitter API. Specifically, we discuss the data related to tweet text, Twitter users and other metadata which exist in Twitter JSON objects. Then, we define and discuss various textual features, image and video features, Twitter metadata-related features and network features that can be extracted from them. We also discuss applications that use different feature types along with a justification for why certain features perform well in the context of informal short text messages such as tweets. Then, we discuss five Twitter-based applications that utilize the different feature types and highlight the features that perform well in the corresponding application setting. Finally, we conclude the chapter by discussing Twitris, a real-time semantic social web analytics platform that has already been commercialized, and its use of Twitter features.

[1] http://www.pewinternet.org/fact-sheet/social-media/
[2] https://www.omnicoreagency.com/twitter-statistics/

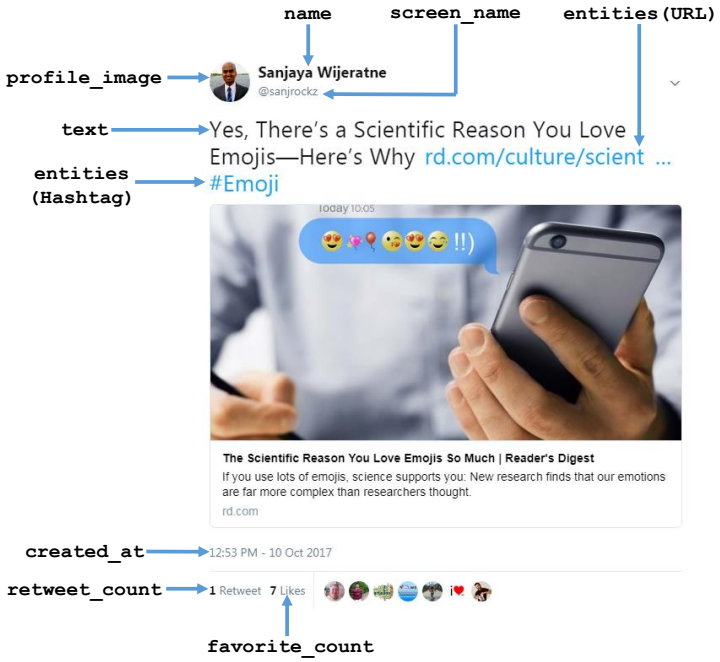

FIGURE 1.1: An example tweet.

1.2 Data Present in a Tweet

Twitter provides two main APIs, namely, the Streaming API[3] and the REST API[4] to access the tweets published by the Twitter users in real time. Application developers can use Twitter Streaming API to collect a random sample of tweets, or a sample of tweets that match a particular keyword, or a set of tweets that originate from a particular geographical location. Twitter REST API can be used to collect a user's past tweets, friends and followers, and user profile information among many other attributes. Both APIs provide the facility to download Twitter data in the JSON format. We analyze the different data types present in a tweet JSON object (shown in Figure 1.1), with a special focus on the data related to the tweeted text, the user who authored the tweet, location names and URLs in the tweet text, etc.

[3]https://dev.twitter.com/streaming/overview
[4]https://dev.twitter.com/rest/public

1.2.1 Tweet Text-related Data

This section discusses the textual data available in a Twitter JSON object. Data fields corresponding to tweet text that are commonly used in Twitter-based applications are discussed below with pointers to applications that use them.

1. `text`: `text` field in the Twitter JSON object holds the textual content present in a tweet. Tweet text and features derived from it (e.g., part-of-speech tags of the words) are extensively used in Twitter-based applications [14, 79, 4, 28, 91].

2. `entities`: This field contains *hashtags, URLs* and *user_mentions* present in a tweet. *Hashtags* were introduced as a way of organizing Twitter conversations by topics, thus, they are extensively used for collecting topic-specific tweets. In addition to that, they have also been used in various Twitter content analysis studies, including information propagation [76]. *URLs* can also contain valuable information to enrich the content, and thus, used in studies related to Twitter content analysis [28]. *User_ mentions* contain all Twitter usernames mentioned in a tweet. User mentions can be helpful in building user interaction networks to find influential users [62].

3. `retweeted`: This field indicates whether the current tweet is a retweet of a previous tweet.

4. `retweeted_status`: If the current tweet is a retweet, then this field contains all the information about the original tweet which was retweeted, including the text of the tweet and information about the user who originally tweeted it. The retweets of retweets are not captured in this field, but only the original tweet. Data present in the `retweeted_status` field are commonly used in studies that analyze tweet text [8], retweet networks [56, 50], influence and information propagation on Twitter [10]. Past research has also shown that having hashtags and URLs in a tweet make it more likely to be retweeted by others [73].

5. `retweet_count`: This field indicates the number of times that the current tweet has been retweeted. Retweet count can be helpful to predict the popularity of a tweet [9].

6. `quoted_status`: This field carries information about the quoted tweets. Quoted tweets are a special form of retweets where a Twitter user gets to write his/her own tweet/message while retweeting a tweet. Information present in a quoted tweet can also be used in applications that analyze twitter content(e.g., understanding political discourse [30]).

7. `favorite_count`: This field records how many people have liked the current tweet. This information has been used as a feature for predicting the popularity of a tweet [9].

8. `in_reply_to_screen_name`: If the current tweet is a reply to another tweet, this field holds the screen name of the user to whom the current tweet responded. This field can be helpful in building user interaction networks to find influential users [62].

9. `place`: This field will carry information about the places associated with a tweet. Users choose a place to be attached to a tweet from a list of pre-defined places. Thus, this does not necessarily represent the place from where the tweet is originated.

1.2.2 Twitter User-related Data

This section discusses the data available in a Twitter JSON object that corresponds to the user who authored the tweet.

1. `description`: This field records the user-provided description of a Twitter profile. In general, Twitter users use the description of their Twitter profiles to provide a short biography about themselves. Thus, applications that classify Twitter profiles use the content extracted from this field to build classifiers [4, 82, 62, 91].

2. `screen_name`: This field holds the user's screen name. A screen name or a Twitter handle is a unique name that a Twitter user selects to identify him/herself on Twitter. This name starts with an '@' symbol and can go up to 15 characters. `screen_name` is used to uniquely identify Twitter users.

3. `name`: This field holds the Twitter user's name. This field can act as a label to refer to users but the same `name` can be shared by many Twitter users, thus, this field should not be used as a unique way to refer to Twitter users.

4. `location`: This field holds a Twitter user-provided location. Twitter users can enter any value as their location, thus, this field will not always carry a valid physical location of a Twitter user. For example, some invalid physical locations can include 'worldwide', 'someone's heart', 'in the middle of nowhere' and 'here'. However, whenever possible, Twitter applications that try to predict the location of a Twitter user utilize the text specified in this field when trying to arrive at a meaningful physical location of a user [71].

5. `geo_enabled`: This field records whether the user has decided to share the geo-coordinates of the tweet's originating location with the tweet or not.

6. `profile_image_url`: This field records the profile image URL used in a given Twitter profile.

7. `profile_banner_url`: This field records the cover image URL used in a given Twitter profile.
8. `followers_count`: This field indicates the number of followers that a given Twitter user has.
9. `friends_count`: This field indicates the number of friends that a given Twitter user has.

1.2.3 Other Metadata

This section discusses the metadata available in a Twitter JSON object that corresponds to either tweet text or the user who tweeted it.

1. `id`: `id` field in the Twitter JSON object represents the unique ID associated with a tweet. Twitter applications can use this ID to uniquely identify and retrieve tweets.
2. `coordinates`: `coordinates` field contains the longitude and the latitude of a tweet's location which identifies the exact location of the user at the time of authoring the tweet.
3. `lang`: This field holds the machine-detected language of a tweet text. Note that this does not necessarily represent the language used in the Twitter profile.
4. `created_at`: This field records the Coordinated Universal Time (UTC) of the tweet creation. This field can be helpful for temporal trend analysis and for applications that try to predict a Twitter user's location.
5. `time_zone`: This field records the time zone associated with a Twitter user profile. A Twitter user can specify the time zone information at the time of their Twitter profile creation.

1.3 Common Types of Features used in Twitter-based Applications

This section presents a list of features that are commonly used in Twitter-based applications. It also discusses how the features are defined and why they are important for those applications. The features are categorized into several groups based on what type of content is present in them such as text, image, and video (examples are given in Table 1.1 for selected feature types). This section also emphasizes why those features tend to perform well with certain types of Twitter applications.

TABLE 1.1: Different feature types extracted from the example tweet shown in Figure 1.1.

Feature Type	Example
Unigrams	Yes, There, 's, a, Scientific, Reason, You, Love, Emojis, Here, 's, Why
Bigrams	Yes There, There 's, 's a, a Scientific, Scientific Reason, Reason You, You Love, Love Emojis, Emojis Here, Here 's, 's Why
Part-of-Speech	Yes/UH, There/EX, 's/VBZ, a/DT, Scientific/NNP, Reason/NNP, You/PRP, Love/VBD, Emojis/NNP, Here/RB, 's/VBZ, Why/WRB
Named Entities	Emojis
Hashtags	#Emoji
URLs	https://www.rd.com/culture/scientific-reason-love-emojis
Image Tags	Male, Person
Creation Time	12:53 PM - 10 Oct 2017
Language	English

1.3.1 Textual Features

This section discusses textual features that can be extracted from Twitter data. As discussed in Section 1.2, a Twitter JSON object contains many fields that hold information in the form of text such as tweet text and profile description of a Twitter user. Next, we define and discuss different types of features that can be extracted from the textual content present in a tweet.

1. **Word n-grams**: Word n-grams are defined as contiguous sequences of n words that appear in a text fragment. Popular n-grams are unigrams containing one word and bigrams containing two words. For example, if we consider the tweet text ``I am feeling happy today'', the set of all unigrams extracted from the tweet text include the words {`I, am, feeling, happy, today''`} while the set of all bigrams extracted from the same tweet include {`I am, am feeling, feeling happy, happy today`}. Word n-grams are very popular in many Twitter-based applications that analyze textual features, including applications that are designed for sentiment analysis [14, 28], emotion analysis [79], and user profile classification [4, 82, 65]. Past research on sentiment analysis has shown that unigram features perform best when Twitter data is used as training data [11], whereas, bigram features outperform unigrams when longer text fragments such as product reviews are used for training [77].

2. **Part-of-Speech (PoS) Tags**: In linguistics, Part-of-Speech (PoS) is defined as categories of words that exhibit similar properties or functions based on how words are used in the language. PoS tags are commonly used in many natural language processing applications including lan-

guage parsing and word sense disambiguation [36]. For example, if we consider the tweet text ‘‘`I'm feeling happy today`’’, a PoS tagger would categorizes the word `I` as a personal pronoun (`I/PRP`), `'m` as a verb which is non-3rd person singular present (`'m/VBP`), `feeling` as a verb, which is gerund or present participle (`feeling/VBG`), `happy` as an adjective (`happy/JJ`), and `today` as a singular or mass noun (`today/NN`). PoS tags are widely used in Twitter-based sentiment and emotion analysis applications and have shown to improve the performance of the baseline sentiment analysis models that are based on word n-gram features [14, 28, 79]. Prior research has shown that natural language processing (NLP) tools that are trained on well-formed text corpora might not work well with social media text due to language variations [67, 84]. Therefore, social media-specific Part-of-Speech (PoS) tagging software has been used in Twitter-based applications [32].

3. **Named Entities**: Named entities are real word objects such as persons, places or organizations that are identified by proper names. Named entities can act as important features for traditional applications such as information retrieval and search, as well as Twitter-based applications such as target-specific sentiment analysis [14]. Named entities are usually extracted from the tweet text [14], profile descriptions [62], and user location [71]. Similar to PoS tag extraction from tweets, named entity extraction from tweets also require specially designed tools due to the informal nature of the language used in Twitter [67, 24]. For example, Twitter users prefer abbreviations and unconventional shortened versions of person names in tweets (e.g., ‘0bama’ and ‘OBMA’ to refer to ‘Barack Obama’) due to the 140 character limitation imposed by Twitter[5]. Thus, many features that are used in named entity extraction tasks in well-formed text such as capitalization of the first letters of words, and punctuation would not work in Twitter-settings. This has resulted in building Twitter-specific named entity recognition tools [67]. Cultural entities, which are a special form of named entities that refer to artifacts of culture such as music albums, movie names, and book names are also common among Twitter conversations [49]. Those could also play an important role in Twitter applications that are targeted for fans of the music artists, movies or books.

4. **Implicit Entities**: Implicit entities are the “entities that are mentioned in text without an explicit mention of their names nor synonyms/aliases/abbreviations or co-references in the same text” [58]. Implicit entities are a common occurrence. For example, past research has found that nearly 21% of movie mentions and 40% of book mentions are implicit in tweets [59]. For example, consider the tweet ‘‘`Aren't we gonna talk about how ridiculous the new`

[5]Twitter is planning to support up to 280 characters for selected languages in near future.

`space movie of Sandra Bullock is?''`. It contains an implicit reference to the movie 'Gravity'. Past research has shown that external knowledge bases such as DBpedia and Wikipedia can be successfully utilized to identify implicit entity mentions in clinical narratives [58] and tweets [59].

5. **Emoticons**: Emoticons are the pictorial representations of facial expressions (e.g., :-), :-D, :-(etc.) composed mainly using punctuation marks and letters. They are commonly used to express the emotion of the message sender or to convey other non-verbal cues [54]. Emoticons are commonly treated as sentiment bearing terms in sentiment analysis applications and emotion bearing terms in emotion analysis applications. They have shown to improve the performance of sentiment and emotion analysis on short text such as tweets [79].

6. **Emoji**: With the rise of social media, pictographs, commonly referred to as 'emoji' have become one of the world's fastest-growing forms of communication[6]. Users of social media platforms use emoji to express their emotions and other non-verbal cues, making them important features for applications that analyze sentiment and emotion [53]. Similar to emoticons, emoji are rarely seen in well-formed, lengthy texts. Thus, they can act as effective features in short text processing applications that utilize them for brevity. Past research on sentiment and emotion analysis on tweets has reported that the accuracy of the two tasks can be improved by using emoji as features [53, 87]. These studies further state that the sentiment polarity of emoji increases with the distance of their appearance from the start of the text in a sentence. Recent research has shown that emoji can also act as an indicator of sarcasm [29]. Many Twitter-based studies convert emoji into their text equivalents (i.e., using 'face_with_tears_of_joy' as a feature instead of processing the image of the emoji) using third-party software [4, 82].

7. **Hashtags**: A hashtag is a special kind of word that starts with the symbol '#'. They were introduced as a way of organizing conversations by topics on Twitter (e.g., #HurricaneSandy hashtag groups all tweets posted on Hurricane Sandy) and are much more commonly seen on social media platforms compared to in grammatical text such as news articles. Hashtags can be essential elements for information retrieval and search from social media sites. They have been extensively used in many real-world applications including disaster management and relief coordination [63], and information propagation [76].

8. **User Mentions**: A user mention is a mention of a username in a tweet, which starts with the special character '@' followed by the username of the Twitter user (i.e., @username). User mentions are used on Twitter

[6] https://goo.gl/jbeRYW

to refer to another Twitter user, whereas, real names are used in grammatical text. User mentions are often used as binary features in social media text analysis applications (i.e., existence or non-existence of a @username in a tweet is regarded as a feature). Social network analysis applications use user mentions to build the user networks [62].

9. **URLs**: A Uniform Resource Locator or a URL is a unique address that locates a resource (e.g., a Web page) on a computer network (e.g., the World Wide Web). Since many users use social media websites to read and share news and other interesting web resources they find online, sharing URL links is a common practice in social media websites including Twitter. Twitter users mainly share links to other web resources in their profile descriptions and tweets. Twitter-based content processing applications such as sentiment analysis applications often treat the existence of a URL in a tweet as a binary feature [14] whereas other content processing application may process the information present in the resource linked to the URL as well [28].

10. **Repeated Letters and Punctuations**: Twitter users often use multiple repeating letters in a word as a way of conveying and emphasizing the emotion or sentiment associated with a tweet. For example, the tweet ``I looooved the movie'' expresses a positive sentiment about a movie viewing experience of a Twitter user. Thus, repetition of letters and punctuation marks have been used as features for Twitter content analysis. These informal language styles are rarely seen in well-formed text, thus, these features do not play a special role in grammatical text analysis.

1.3.2 Image and Video Features

This section discusses image and video features that can be extracted from Twitter posts and user profile data. Twitter users can upload images to their Twitter profiles using three main ways, namely, (i) uploading an image as the profile image, (ii) uploading an image as the cover image, and (iii) uploading an image as part of a tweet. The only option to post a video to Twitter is through a tweet. Twitter API disseminates image and video URLs with Twitter JSON objects if images or videos are present in a Twitter profile or in a Twitter message. Below, we define two main types of features that can be extracted from image and video content present in a tweet or a Twitter profile.

1. **Image Tags**: Image tags are the labels or names associated with images. These names or labels are usually assigned by human annotators so that the annotations along with images can be later used to build image classification models. For example, an image of a car could be associated

with a label 'car', which will then be used to train a machine learning classifier that can automatically tag images of cars. Images extracted from Twitter profiles are used as features in studies that attempt to predict the gender of a Twitter user [68] or Twitter user groups [4, 82].

2. **Video Titles and Text in Video Comments**: Twitter users also post videos along with their tweets and those can provide additional information for tasks at hand. For example, past work on gang member Twitter profile identification suggest us that textual features extracted from YouTube video titles and comments can act as useful features to identify gang member profiles [4, 82].

1.3.3 Twitter Metadata-related Features

This section discusses features that can be extracted from various metadata related to a tweet. Twitter metadata provide useful information that explain different events related to a tweet such as the tweet originating time, tweet originating location, and tweet language. Below, we discuss several metadata-related features that are commonly used in Twitter-based applications.

1. **Geo-coordinates of a Tweet**: Geo-coordinates of a tweet is a combination of the longitude and the latitude of the tweet's originating location. This feature will be available in Twitter JSON object only if the Twitter user who posted the tweet had already agreed to share the location of the tweet's originating location in the tweet. Geo-coordinates of a tweet is very valuable for any Twitter-based application that requires the location of a Twitter user or a tweet as they can provide the exact tweet originating location. Thus, geo-coordinates are commonly used in location-based Twitter applications [71, 27].

2. **Tweet Creation Time**: Tweet creation time records the Coordinated Universal Time (UTC) of the tweet creation. This can be a helpful feature to organize a collection of tweets chronologically for further processing. For example, dialog processing systems that process tweets can be benefited from this feature to organize the messages exchanged among a group of users before further processing them to glean contextual information related to the discussion among Twitter users.

3. **Language of a Tweet**: Twitter also provides the machine-detected language of a tweet. This can be a helpful feature for information filtering when collecting tweets using a given language. This could also be helpful for Twitter-based user demography studies.

1.3.4 Network Features

This section discusses features that can be extracted from various Twitter networks. We define network-based features as the features extracted from Twitter user network and the metadata associated with them. A network (graph) is a set of nodes (vertices) connected by a set of edges (arcs) where the connections among nodes symbolize a relationship among them. For example, a simple Twitter-based network could be a graph of a user's friends on Twitter where Twitter users are represented by the nodes in the graph and the friendship among users are represented using the edges. There are two main types of networks available in Twitter, namely, user interaction-based networks and friend-follower networks [39]. Different types of relationships among Twitter users are used to generate Twitter networks, thus, they can yield different types of features based on the relationships used to generate the networks.

1. **Retweet**: Retweet is the task of sharing an existing tweet so that the tweet would reach the followers of the user who shares the tweet. This has become a very common practice among many Twitter users [50]. Retweets can be used to identify influential users [10, 3, 62]. For example, retweet pattern among a set of Twitter users can be used to derive features that can help identify influential users and how influence propagates.

2. **Reply**: A reply is a tweet that is targeted to a Twitter user in response to a previous tweet posted by the user. Replies typically start with a '@username' followed by the message posted as the reply. Replies are commonly used to construct the interaction networks among Twitter users. Twitter reply-based networks can be used to derive additional features [89, 92], that can be helpful in deciding the influential users and clustering Twitter users based on their interests.

3. **Followers and Friends**: A Twitter follower is a Twitter user that a particular Twitter user follows. If two Twitter users follow each other, they are called friends. Friends and followers networks have been used in many applications for the identification of a Twitter user's social ties, a user's influence on others and information propagation.

1.4 Twitter Feature Engineering in Selected Twitter-based Studies

This section discusses how different features extracted from Twitter can be used in real-world applications. Specifically, this section discusses studies

ranging from Twitter user profile classification and prediction to gleaning the sentiment and emotion of Twitter users.

1.4.1 Twitter User Profile Classification

Twitter user profile classification is a well-studied problem where a class label is assigned to a Twitter profile from a set of pre-defined labels. Concrete examples of Twitter profile classification include user political affiliation classification [57], ethnicity classification [57], gender identification [42], brand loyalty prediction [57], and user occupation classification [65]. Majority of these applications rely only on textual features extracted from content posted on Twitter or user profiles. Balasuriya *et al.* recently studied how to identify the Twitter profiles of street gang members where they used different types of content-based features and user-based features [4, 82]. This problem was motivated by the recent increase in the number of gang violence-involved homicides which were traced back to the fights that first started on online social media [86].

Balasuriya *et al.* first curated a large dataset of gang member profiles on Twitter that consists of 400 authentic gang member profiles and 2,865 non-gang member profiles [4]. For each Twitter profile in their dataset, they collected up to 3,200 most recent tweets that were associated with those profiles along with profile descriptions and images (profile and cover photos) of every gang and non-gang member profile. They analyzed the text in the tweets and profile descriptions (unigrams) of those Twitter profiles, emoji use, profile images, and music interests. In their analysis, they found that 5.72% of all words posted by gang members were curse words, which is nearly five times more than the average curse word usage on Twitter among the general population [80]. They also noticed that the gang members often talk about drugs and money with terms such as *smoke, high, hit*, and *money*, while ordinary users hardly speak about finances and drugs. Ordinary users often vocalize their feelings with terms such as *new, like, love, know, want, look, make, us*. These differences give a clear indication that the words used by gang and non-gang members could act as important features for gang member profile classification. They found that the terms *rip* and *free* appear in approximately 12% of all gang member Twitter profile descriptions, which suggests that gang members use their profile descriptions as a space to grieve for their fallen or incarcerated gang members. They also found that 51.25% of the gang members in their dataset had a tweet that linked to a YouTube video and 76.58% of those tweets were related to gangster rap music. Emoji analysis revealed that the fuel pump emoji was the most frequently used emoji by the gang members, where it was often used in the context of selling or consuming marijuana. The pistol emoji was the second most frequent emoji, which was often used with the police cop emoji in an 'emoji chain' to reflect anger at police officers. They reported that 32.25% of gang members had chained together the police and the pistol emoji, compared to just 1.14% of non-gang mem-

bers. Gang members were often seen holding or pointing weapons in groups displaying gangster culture, or showing off graffiti, hand signs, tattoos, and bulk cash in their profile or cover pictures.

Balasuriya *et al.* used four different classification models for their task: a Naive Bayes net (NB), a Logistic Regression (LR), a Random Forest (RF), and a Support Vector Machine (SVM). These four models were chosen because they are known to perform well over text features, which is the dominant type of feature considered. They used 10-fold cross validation for training and evaluating the classifier models. First, they trained a series of classifier models by using features extracted from a single feature type (i.e., tweet text, emoji, profile, image, or YouTube) as training data. The classifiers that use a single feature type were intended to help them study the quality of the predictive power of those features alone. Then they also trained another series of classifier models that consider all types of features combined. Their results reported that the RF classifier model trained on the unigram features extracted from tweets perform reasonably well, with a $F1$-score of 0.72 for the 'gang' class. NB classifiers trained on the unigram features extracted from YouTube videos and emoji shared in tweets were the next best classifier models with $F1$-scores 0.65 and 0.61, respectively, for the 'gang' class. They reported that the performance of the classifier models improved when different types of features were combined ($F1$–score of 0.77 for RF classifier). In a later study, Wijeratne *et al.* improved the above classification models ($F1$–score of 0.78) using word embeddings [82].

1.4.2 Assisting Coordination During Crisis Events

Social media, specifically, Twitter has proven to be an effective medium for sharing information during crisis events such as Hurricane Sandy and Typhoon Yolanda. The shared information often includes important messages such as requests for help (seekers) as well as responses (suppliers) to such requests. However, extracting seeker and supplier information can be challenging as crisis events tend to produce large amounts of data within a short period of time [64]. Thus, locating tweets with seeker and supplier information becomes a high-priority task for the disaster coordination teams who monitor Twitter during disaster events. To assist online disaster coordination teams with extracting supplier and seeker information, Purohit *et al.* studied the problem of matching resource requests (seekers) with responses to help (suppliers) as an intent classification problem [64] where an intent is defined as a purposeful action. Intent mining on social media can be challenging, specially, in the context of disaster relief coordination due to numerous reasons, including but not limited to:

1. Informal language use causes ambiguity in interpreting user expressions in short-text messages, weakening predictor-class relationships (e.g.,

'wanna help' appears as a strong intent signal but exists in messages of two complementary intent classes, 'seeking' and 'offering').

2. The sparsity of instances of specific intent classes in the corpus creates data imbalance (e.g., [63] showed that expressions of 'offering' intent were only a fraction of those with 'requesting' intent (1:7 ratio) during Hurricane Sandy event in 2012).

3. Both intent of 'seeking' and 'offering' may co-occur within a single message. The limited motives pertain to the transactional intent of buying and selling, which are different from the critical actions involved in our problem context of cooperation.

TABLE 1.2: Examples of short-text documents and potential intents.

Short-text Document	Intent
What is the location of the nearest #Redcross? I wanna give blood to help victims of the Hurricane	Offering Help
@Zurora wants to help @Network4Good with Hurricane relief. TEXT Sandy 8088 & donate $10 to @redcross	Seeking Help
Would like to urge all the citizens to make proper preparations for the Hurricane	Advising
Thx to all in Dayton who brought supplies for those who affected by Hurricane Sandy.	Acknowledging

Table 1.2 exemplifies a few tweets which show seeker/supplier behavior. Past research has investigated simple pattern-based approaches to identify seekers and suppliers and reported that pattern-based methods alone cannot address issues such as ambiguity in the informal language used in seeker/supplier tweets [63]. Thus, Purohit *et al.* have used machine learning models with three different types of features extracted from seeker/supplier related tweets and expert knowledge (e.g., features extracted from common messages patterns in a disaster-related event), which are discussed below.

1. Bottom-up: They are unigram features extracted from the tweet text.

2. Top-down: Features based on knowledge guided patterns, social knowledge guided patterns, and contrast mining guided patterns. Knowledge-guided patterns rely on semantic-syntactic knowledge of intent expression. For example, a subject with the main verb 'have' and any noun suggests an offering intent. However, the same text preceded by the auxiliary verb 'do' and the pronoun 'you' suggests a seeking intent. Similarly, word order such as verb-subject position also plays a crucial role in intent expression. Such patterns for expressing intent can help address the ambiguity challenge by endorsing the likelihood of an intent association for a short-text document. The pattern design leverages a lexicon of verbs,

given that verbs imply a plan for action. Using Schanks P-Trans primitive [69], which reflects the transfer of property, Purohit *et al.* acquire seeking-offering intent related verb classes. Their verb lexicon includes the Levin verb [41] categories of {give, future, having, send, slide, carry, sending/carrying, put, removing, exerting force, change of possession, hold/keep, contact, combining/attaching, creation/transformation, perception, communication}. Their patterns also include classes of auxiliary verbs (e.g., be, do, have), the models (e.g., can, could, may, might), question words ('wh'-words and how), and the conditional (if). They extend the seed patterns by an exhaustive representation of synonymous verbs preserving the tense, using the WordNet knowledge base [47]. Each pattern is treated as a binary feature. Social conversation can help further differentiate the intent, especially, during a crisis event. During crisis events, citizens usually get involved in conversations regarding resource seeking and supply. As a part of this, stop words, which are often discarded, are considered as a conversation indicator for the social knowledge guided patterns. The full list of the social knowledge guided patterns used in this study is available in [64]. To learn contrast mining guided patterns, Purohit *et al.* used a state-of-the-art emerging pattern detection algorithm [25, 43], which is capable of identifying patterns in tweet text that are dominant in one class (e.g., seeker) but not often seen in the other class (e.g., supplier).

3. Hybrid: In the hybrid feature selection, both the top down and bottom up features were combined and used in the classification process.

Purohit *et al.* used two datasets for their experiments. The first dataset (Dataset 1) consisted of 4.9 million tweets for the event of Hurricane Sandy in the U.S. in 2012 and the second dataset (Dataset 2) consisted of 1.9 million tweets for the event of Typhoon Yolanda in the Philippines in 2013. Their training dataset consisted of 3,135 tweets extracted from Dataset 1 and 2,000 tweets extracted from Dataset 2. Three human judges were asked to annotate each tweet in the training dataset using six labels {Request to get, Offer to give, Both request and offer, Report of past donations, None of the above, Cannot judge} for ground truth data creation. They merged the labels to design the intent classes {seeking, offering, none (i.e. neither seeking nor offering)}. They used Random Forest algorithm (RF) to train classifier models and used 10-fold cross-validation to evaluate them. They reported that the classifier models trained using the hybrid features outperformed the classifier models trained on the top-down and bottom-up features alone. The absolute gain in the $F1$-score and accuracy for the hybrid approach are up to 7% and 6% respectively, compared to the next best performing classifier models trained using top-down and bottom-up features [64]. They also reported that more than half of the most discriminating, best ranked 1% features belong to the top-down (knowledge-based) processing representation, which shows the

importance of encoding world knowledge into improving the machine learning classifiers [70].

1.4.3 Location Extraction from Tweets

Location extraction or location identification problem in Twitter is an important problem where the focus can be either to identify the Twitter user's location or the tweet's originating location or both. Knowing the location of a Twitter user or tweet origination can be beneficial in many applications. For example, Carmen *et al.* [27], extract location names from tweets for a surveillance application of influenza by tapping on the spatial extents of tweets and their content. A wide range of applications also use the location information extracted from tweets such as to study U.S. elections [14], users' religiosity [15], analysis of drug use across geographies in the USA [20, 19], the effect of governmental decisions on the legalization of drugs and the perception of Twitter users on such decisions [40], and even cross-cultural differences and their effect on the mental illness of Twitter users [23]. All of the above studies and use-cases only need coarse level location information (such as country, state, or neighborhood levels) as opposed to other studies which require finer levels of location information (such as neighborhood, street, or building). Crisis-related applications have critical need for fine-grained location information. For example, to study the inter-dependencies between resources and needs during a disaster, Bhatt *et al.* [6] exploited location mentions in text and other information about the resource needs to better facilitate disaster relief efforts. Anantharam *et al.* [2] used a knowledge base of location names (gazetteer) to help in extracting spatiotemporal proximity of city events such as traffic and road closures due to accidents. State of the art social media data processing systems, such as Twitris [71], support spatio-temporal-thematic analysis of events on social media with real-time monitoring. The first item of the analysis triple is the spatial context of each event, allowing users to filter those events based on their locations. There are four types of location information available on Twitter:

1. A Twitter user can specify his/her location in the user's profile and this can be retrieved from the `location` field in a Twitter JSON object.

2. A Twitter user can share the tweet's originating location (longitude and latitude) along with the Tweet. This information can be retrieved from the `coordinates` field in a Twitter JSON object.

3. A Twitter user can choose a location/place from a list of pre-defined places to be attached to tweets. This information, which consists of the place name and the place bounding box, can be retrieved from the `place` field in a Twitter JSON object.

4. A Twitter user can include location names in the tweet. There's no

guarantee that the place names extracted from tweet text is Twitter user's or tweet's originating location, however, prior research has shown that the location names [38] and linguistic clues [16] extracted from tweet text can be used to improve Twitter user location prediction.

The exploitation of the geo-coordinates (longitude and latitude) of a tweet becomes a straightforward task if the Twitter user authorizes Twitter to share the tweet's originating location. However, since only around 2% of the tweets contain such information, studies tend to exploit other means to infer the location of a tweet or a Twitter user. This is a non-trivial problem. Finding the user profile location or extracting the tweet originating location from tweet content might not result in the actual tweet's originating location for many reasons. For example, Twitter users might be traveling, thus, their tweets could be authored from different locations than what is specified in their Twitter profiles. Similarly, the tweet content might be about a remote location that the user is interested in but does not reflect the users actual location or the tweet originating location. Studies have adopted different features extracted from tweets, and user profiles to estimate the location of a Twitter user or a tweet. Carmen *et al.* [27] infer the location of Twitter users and tweets from the user's profile location specified in a Twitter profile. After identifying the irregularities in the Twitter naming scheme in specifying the location names, they used a list of location name synonyms to map lexical variants of a location name to a common location name. They also developed methods to normalize location name acronyms such as "NYC" to "New York City" and filtered out location names such as "Neverland". Finally, they used Yahoo's PlaceFinder API[7] to find the geo-coordinates of the remaining location names. Exploiting the information from the users and their follower/followee network data, the state-of-the-art geo-localization tool Pigeo [66] finds the location of users given their textual messages and their interactions with other users. The tool uses pre-trained geo-location models to predict the user geolocation and achieves on average around 65% accuracy. Notably, Mourad *et al.* [48] found that the language (and its coverage) of a Twitter user highly influences the accuracy of geolocating the user.

Twitris [71], a real-time social media event monitoring tool computes the location of a tweet using the tweet location and the user profile location discussed earlier. For example, if a tweet is not associated with exact geo-coordinates of the tweet originating location, Twitris tries to utilize the user's profile information. This two-step process results in assigning tweet originating locations up to 25% of tweets. Extracting location names from the tweet text is the most difficult task among the four types of location extraction methods discussed earlier. Similar to other NER tasks, extracting location names starts with delimiting the location names in text (i.e., identifying the mention of the word "London" as a possible location name), linking them with a gazetteer or a knowledge base of location names and disambiguating the

[7] https://developer.yahoo.com/geo/placefinder/

identified location names (i.e., deciding whether the mention of London refers to "London, Ohio" or "London, England"). Prior research has shown that natural language processing (NLP) tools trained on well-formed text corpora might not work well with social media text due to language variations [67, 84]. Therefore, social media-specific NLP methods have been developed for NER and Part-of-Speech (PoS) tagging, which is commonly used in NER to improve the performance [7, 67, 31]. Other techniques for NER include phrase mining or noun-phrase extraction, and n-gram matching [44, 31, 46]. Methods such as [46] pre-loads a list of location names of an area of interest to improve the accuracy of the location name extraction process. Additionally, methods such as [35] have devised semi-supervised methods which use beam-search and structured perceptron to extract and link location names to their respective Foursquare records. The state-of-the-art methods such as [1] extract location names from tweets and link them to open street maps gazetteers using statistical language models. The method discussed in [1] achieved a combined $F1$-Score of 81% by avoiding the use of unreliable syntactic and orthographic features of tweets.

1.4.4 Studying the Mental Health Conditions of Depressed Twitter Users

With the rise of social media, millions of people are routinely expressing their moods, feelings, and daily struggles with mental health issues on social media platforms like Twitter. Unlike traditional observational cohort studies conducted through questionnaires and self-reported surveys, mining social media for depressive symptoms expressed in tweets provide a new way to capture the behavioral attributes such as mood and day-to-day activities that are related to one's thinking patterns. Thus, much progress has been made in studying mood and mental health conditions through the lens of social media, recently [90, 52, 22, 91, 5, 37]. The vast collection of mental health studies on social media can be grouped into two major categories, namely, lexicon-based [37, 51] and supervised modelling [52, 22, 21]. These studies suggest that the people's language style, emotion, ego-network, and user engagement can be used as discriminating features to recognize the depression-indicative posts. In particular, these distinguishing characteristics can be used to build a model to predict the likelihood of depression expressed in a post [22] or in an individual [21]. In another study, Nguyen *et al.* [52] employ affective aspect, linguistic style and topics as features for characterizing depressed communities. LIWC[8] program has been employed to capture the linguistic style signals, including auxiliary verbs, conjunctions, adverbs, functional words, and prepositions as well as the number of positive and negative words. Wang *et al.* [81] utilizes various features including sentence polarity for the detection of depressed individuals in Twitter. By training a text classifier that identifies

[8] http://liwc.wpengine.com/

three different mood states (anxiety, anger, and depression), Reis *et al.* [26] studies the effect of exercise on mental health. They use n-grams extracted from tweet text, psychological categories, and emoticons as features. Similarly, Shuai *et al.* [72] builds a model for identifying potential cases of social network mental disorders.

Moreover, Coppersmith *et al.* [17] created a dataset of 1,800 Twitter profiles that can be used to develop methods to automatically identify the depressed users on Twitter. A system that used topic modeling and bag-of-words extracted from tweet text, LIWC features, metadata and clustering features achieved state-of-the-art performance [60] when applied to the above dataset of 1,800 depressed Twitter users. Another related line of research which is similar to identifying depressed users on Twitter is studying suicide and self-harm signals from Twitter posts [34, 74, 33]. By studying the tweets posted by individuals attempting to commit suicide, Coppersmith *et al.* [18] reports quantifiable signals of suicide ideations. On the other hand, in the context of lexicon-based approaches, Karmen *et al.* [37] leverages a dictionary-based approach for assigning an overall depression score to each subject. They simply count all the existing phrases that are matched with depression indicators. In another related study, Park *et al.* [55] reports that the usage of keyword "depression" in tweets. They show that individuals tweet about their depression and even disclose updates about their mental health treatment on Twitter. They found an association between excessive use of negative emotions-related words and having major depressive disorder (MDD). In contrast, no relation has been found in the use of positive emotions-related words and depression.

Similarly, Neuman *et al.* [51] reports a natural language processing approach for identification of depression by utilizing a depression lexicon incorporating both metaphorical and non-metaphorical words and phrases. They perform a web search to seek documents containing text patterns such as "depression is like *", where "*" can be any word. Employing a dependency parser, they extract phrases which could possibly be the sources of depression and expand them by their first and the second-degree synonyms obtained from a depression lexicon used in the study. However, in natural language, words can be ambiguous. For instance, depression may be used to express different concepts such as 'economic depression', 'depression era', 'great depression', and 'tropical depression'. Moreover, neurotypical people use this term to express their transient sadness. An example of a neurotypical user tweeting about depression could be `''I am depressed, I have a final exam tomorrow''`. Furthermore, the experience of depression may be expressed implicitly making lexicon-based approach insufficient for accurate fine-grained analysis of depression symptoms over time. Another inherent drawback of all lexicon-based methods is their high precision at the expense of low recall and lack of context-sensitivity. For example, 'sleep forever' indicate suicidal thoughts rather than sleeping activity.

In contrast with lexicon-based approaches, the supervised approaches require labor-intensive annotation of a large dataset for reliable training. More

recently, Yazdavar *et al.* [91] developed a statistical model for linguistic analysis of social media content authored by a subject by seeking depression indicators and their variation over time. Particularly, they developed a probabilistic topic modeling over user tweets with partial supervision (by leveraging seeded clusters), named semi-supervised topic modeling over time (ssToT), to monitor depression symptoms. Their experimental results revealed that there are significant differences in the discussion topic preferences and word usage patterns in the self-declared depressed user on Twitter against the non-depressed users. Their model showed promising results with an accuracy of 68% and a precision of 72% for capturing depression symptoms per user over a period of time [91].

1.4.5 Sentiment and Emotion Analysis on Twitter

Identifying people's attitude with respect to a specific topic, context, or interaction, and analyzing subjective experiences have fueled research interest in sentiment analysis/opinion mining [14, 28, 79, 78]. Many opinion mining models and tools have been developed to glean the attitudes of people towards products, people, topics, product attributes (e.g., display screen of a phone: big screen size can be a positive feature for a phone) and aspects (e.g., a feature and a range of values that an attribute can take such as {screen, big, small, medium}) in various contexts. Below, we discuss recent studies on opinion mining and emotion analysis on Twitter.

In opinion mining tasks, a subjective experience (e.g., sentiment, emotion, intent etc.) is defined as a quadruple *(h, s, e, c)*, where h is an individual who holds the experience, s is a stimulus (or target) that elicits the experiences (an entity or an event), e is a set of expressions that are used to describe the experiences (the sentiment words/phrases or the opinion claims), and c is a classification or assessment that characterizes or measures the experiences [11]. In this problem setting, our interest is the user's sentiment or emotion expressed in a tweet, and the opinion holder h is the author of the tweet. More recent methods extract the opinion target s of a subjective experience based on clustering, where product attributes and aspects present in the opinionated text are simultaneously clustered to find the opinion target s [12]. To understand the problem setting of extracting opinion targets, consider the following example product reviews where **explicit** and *implicit* product attributes are shown (in boldface and italics, respectively) along with groups of aspects extracted from the opinionated text below.

1. The phone runs *fast* and *smooth*, and has great **price**.
2. Good **features** for an *inexpensive* android. The **screen** is *big* and *vibrant*. Great **speed** makes *smooth* viewing of TV programs or sports.

In review 1, price is an explicit product attribute, and opinion words "smooth" and "fast" imply implicit product attribute "speed". The task is

to identify both explicit and implicit features, and group them into aspects such as speed, fast, smooth, size, big, price, inexpensive. Given a set of product reviews, Chen *et al.* [12] first use part-of-tagging to identify nouns/noun phrases, verbs and adjectives as candidate product attributes that needs to be clustered. Then they calculate seed terms for each cluster by taking the most frequently appearing candidate words in text. The remaining candidate words are placed to the closest cluster. To group product attributes into aspect groups that are related to a particular domain of interest, they have proposed a novel domain-specific similarity measure incorporating both statistical association and semantic similarity between a pair of candidates attributes. Their algorithm identifies the aspect groups along with the product attributes of each aspect group. In another study, Chen *et al.* have also developed techniques to extract sentiment-related expressions e from opinionated text [14, 13]. They have trained several machine learning classifiers that utilize the information extracted from opinionated text and have successfully used them to predict the outcome of the 2012 presidential election of the USA [14]. Chen *et al.* reported that the textual features extracted from the tweets such as n-grams and part-of-speech tags played a prominent role when building the sentiment analysis classifiers for the U.S. election outcome prediction task.

Emotion analysis is a similar problem to sentiment analysis where opinion holder has expressed his/her opinion as an emotion in the opinionated text. We have also developed automatic methods to extract large collections of labeled data using emotion-related hashtags used in Twitter and trained classifiers to identify seven emotion categories, namely, joy, sadness, anger, love, fear, thankfulness, and surprise [79, 78]. Using 131 emotion-related hashtags, Wang *et al.* collected and filtered out nearly 2.5 million tweets that belongs to the seven emotion categories listed above. Then they analyzed the content-based features extracted from tweet text and used them to train a Multinomial Naive Bayes (MNB) and a LIBLINEAR classifier. Specifically, they looked at unigrams, bigrams, trigrams extracted from tweets, the location of the n-grams (Wang *et al.* hypothesized that the words that are located towards the end of a tweet contribute more towards the emotion expressed in it), emotion words appeared in tweets as identified by multiple sentiment and emotional bearing word dictionaries such as LIWC, MPQA Lexicon[9], and WordNet-Affect and part-of-speech tags of words in tweets. They reported that combining unigrams and bigrams yields better performance than using unigrams alone and that the positional information of the n-grams did not contribute to improving the classification accuracy. They further reported that adding emotion lexicon-based features and part-of-speech features did improve the performance of the classifiers in a small margin. They achieved the best accuracy of MNB classifier of 61.15% and 61.63% of the LIBLINEAR classifier when they combined unigrams, bigrams, emotion lexicon-based features and part-of-speech features. In the above studies, Wang *et al.* did not explore

[9] http://mpqa.cs.pitt.edu/

emoji features. However, emoji are very commonly used in social media and emoji features have shown promising results in improving sentiment [53] and emotion [88] analysis tasks. Specifically, with the recent introduction of emoji sense knowledgebases [83, 84] and the utilization of emoji embeddings for sentiment analysis tasks [85], we have shown that the results of the sentiment analysis could be further improved using emoji features.

1.5 Twitris: A Real-time Social Media Analysis Platform

This section discusses how the Twitter features discussed earlier are used in building a real-time social media analysis platform called Twitris [71]. Twitris is a semantic social web platform that facilitates understanding of social perceptions about real-world events by analyzing user-generated content on social media. It supports analyzing of social media data on multiple dimensions including the Spatio-Temporal-Thematic (STT), People-Content-Network (PCN) and Sentiment-Emotion-Intent (SEI) dimensions. It has been used in analyzing events on Twitter that are related to brand loyalty identification [62], disaster relief and coordination[63, 64, 6], elections and other political phenomena [14, 28], identifying instances of cyberbullying[10], understanding the severity of depression [91, 90], and monitoring marijuana-related chatter [20, 19, 40]. Twitris' major success stories include predicting the outcomes of the 2012[11] and 2016[12] U.S. presidential elections and Brexit[13], among many others. Next, we briefly discuss the different types of features that are being used in STT, PCN and SEI dimensions to demonstrate how feature selection can be incorporated into building comprehensive tweet processing systems.

Spatio-Temporal-Thematic data analysis supports (i) collecting user-generated tweets pertaining to an event from Twitter, along with associated news, multimedia, and Wikipedia content, (ii) processing the collected tweets to extract strong event descriptors such as key phrases, and entities associated with the event to learn the event context, and (iii) presenting summarized data and visualizations including interactive maps. STT processing uses textual features extracted from tweets such as unigrams and bigrams, named entities, hashtags, user mentions, and URLs, to identify key phrases and entities associated with an event (thematic processing). The textual features are further processed using event-specific machine learning classifiers to remove irrelevant data about an event and semantically enhance by linking them with DBpedia knowledgebase. STT dimension also uses Twitter metadata-related

[10] https://goo.gl/A1XYDJ
[11] https://goo.gl/8UR4Ku
[12] https://goo.gl/SctCb7
[13] https://goo.gl/itofGP

features such as geo-coordinates extracted from tweets and profile descriptions to identify a tweet's location (spatial processing). It also uses tweet creation time to analyze tweets generated during a user-selected time period to support temporal queries.

People-Content-Network supports the analysis of the social media users (People), the data shared on social media websites (Content), and the network of social media users (Network) related to an event. It facilitates studying of the information propagation and identification of influential Twitter users in an event-specific user interaction network [62, 65]. For each event that is being monitored, retweets, replies and user mentions in the collected tweets are used as features to create reply and user mention, and retweet networks. Then, different centrality measures such as degree centrality, betweenness centrality, and PageRank centrality are derived as features to identify the influential users. These networks also facilitate analysis of strongly or weakly connected users, which help in assessing coordination and engagement in the network. PCN processing has been extensively used in disaster relief coordination efforts.

Sentiment-Emotion-Intent extraction supports analyzing social media content to extract insights about a user's sentiment related to products, and events [13, 14, 12], mood or emotion related to seven emotions categories [79], and intent behind an action [64, 6]. Twitris supports target-specific sentiment analysis where the sentiment of a tweet is evaluated with respect to a target entity [14, 13]. It identifies seven emotion categories, namely, joy, sadness, anger, love, fear, thankfulness, and surprise in users' Tweets [79]. Sentiment and emotion analysis in Twitris use textual features extracted from tweets such as unigrams and bigrams, part-of-speech tags, named entities, and emoticons. Intent mining on Twitris is mainly supported in the context of disaster relief and coordination. It supports identifying the intents of information seekers and resource donors' (suppliers) and helps to connect them. Intent mining uses textual features such as unigrams, linguistic features and patterns extracted from disaster-related tweets. Section 1.4.2 provides an in-depth discussion on the features used in intent mining in Twitris.

Figure 1.2 depicts selected capabilities of Twitris where the tool is set up to analyze tweets that are related to depression. (1) shows the Twitris control panel to navigate between different widgets dedicated to visualizing different data analysis methods used in Twitris. (2) shows how to select date and range (day, week, month, year), keywords and population density to narrow down data analysis. (3) shows country-level view of depression data, color coded by sentiments expressed in the tweets (Twitris can analyze data for different locations granularities such as country, state, and county level). (4) shows depression data for counties in California, color-coded by sentiments expressed in the tweets. (5) shows the total number of depression-related tweets and the most popular topics discussed on October 6th, 2014 in Los Angeles, CA (also show per capita GDP, unemployment rate) (6) shows a sample of depression related tweets on October 6th, 2014 from Los Angeles, CA. (7) shows

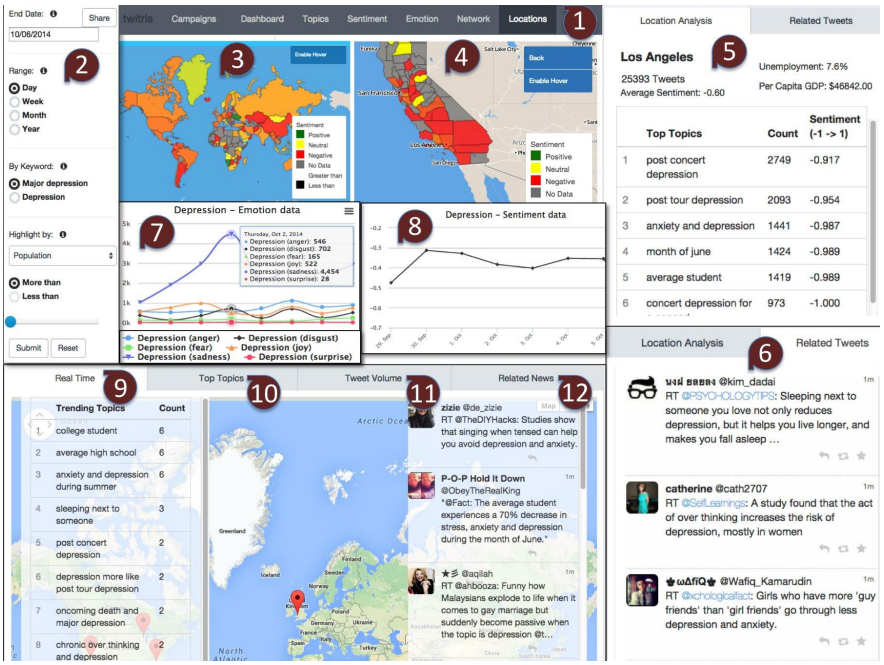

FIGURE 1.2: Data processing and visualization using Twitris

emotion analysis on six emotions (anger, fear, sadness, joy, disgust, surprise) for depression-related tweets collected from September 29th, 2014 to October 5th, 2014. (8) shows sentiment analysis for depression-related tweets collected from September 29th, 2014 to October 5th, 2014. (9) shows trending topics based on the analysis of real-time tweets as of October 6th, 2014. (10) shows the most popular topics related to depression in general – not specific to a location. (11) shows the tweet volume over a period of time (day, week, month), and (12) shows the depression-related news extracted from tweets.

1.6 Conclusion

This chapter discussed feature engineering for Twitter-based applications. In doing so, it first examined the different types of data fields available in a Twitter JSON object. Then it defined different types of features that can be extracted from tweets and Twitter user profiles. It also discussed the Twitter-specific features and why they would perform well with short text. It then discussed five Twitter-based applications where those features are used along

with other computational techniques to solve real-world problems, while highlighting the features that lead to achieving best results in each problem setting. Finally, it concluded the chapter by discussing Twitris, a real-time semantic social web analytics platform that has already been commercialized, and its use of Twitter features. More information about Twitris and the Twitter-based applications covered in the text, along with their corresponding publications, can be found at http://knoesis.org/projects.

1.7 Acknowledgement

We acknowledge partial support from the National Institutes of Health (NIH) award: MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression", the National Science Foundation (NSF) awards: CNS-1513721: "Context-Aware Harassment Detection on Social Media", and EAR 1520870: "Hazards SEES: Social and Physical Sensing Enabled Decision Support for Disaster Management and Response". Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the NIH or NSF.

Bibliography

[1] Hussein S Al-Olimat, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. Location name extraction from targeted text streams using gazetteer-based statistical language models. *arXiv preprint arXiv:1708.03105*, 2017.

[2] Pramod Anantharam, Payam Barnaghi, Krishnaprasad Thirunarayan, and Amit Sheth. Extracting city traffic events from social streams. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(4):43, 2015.

[3] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[4] Lakshika Balasuriya, Sanjaya Wijeratne, Derek Doran, and Amit Sheth. Finding street gang members on twitter. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, volume 8, pages 685–692, San Francisco, CA, USA, August 2016.

[5] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the EACL*, volume 1, pages 152–162, 2017.

[6] Shreyansh P Bhatt, Hemant Purohit, Andrew Hampton, Valerie Shalin, Amit Sheth, and John Flach. Assisting coordination during crisis: a domain ontology based approach to infer resource needs from tweets. In *Proceedings of the 2014 ACM conference on Web science*, pages 297–298. ACM, 2014.

[7] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. In *RANLP*, pages 83–90, 2013.

[8] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences (HICSS)*, pages 1–10. IEEE, 2010.

[9] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Computer Supported Collaborative Work and Social Computing (CSCW)*, pages 211–223. ACM, 2014.

[10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *4th International AAAI Conference on Web and Social Media (ICWSM)*, 10(10-17):30, 2010.

[11] Lu Chen. Mining and analyzing subjective experiences in user generated content. 2016.

[12] Lu Chen, Justin Martineau, Doreen Cheng, and Amit Sheth. Clustering for simultaneous extraction of aspects and features from reviews. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–799. ACL, 2016.

[13] Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *6th International AAAI Conference on Web and Social Media (ICWSM)*, volume 2, pages 50–57, 2012.

[14] Lu Chen, Wenbo Wang, and Amit P Sheth. Are twitter users equal in predicting elections? a study of user groups in predicting 2012 us republican presidential primaries. In *4th International Conference on Social Informatics (SocInfo)*, pages 379–392. Springer, 2012.

[15] Lu Chen, Ingmar Weber, and Adam Okulicz-Kozaryn. U.s. religious landscape on twitter. In *6th International Conference on Social Informatics (SocInfo)*. Springer, 2014.

[16] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[17] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 31–39, 2015.

[18] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–117, 2016.

[19] Raminta Daniulaityte, Robert Carlson, Farahnaz Golroo, Sanjaya Wijeratne, Edward W Boyer, Silvia S Martins, Ramzi W Nahhas, and Amit P Sheth. time for dabs: Analyzing twitter data on butane hash oil use. *Drug & Alcohol Dependence*, 156:e53–e54, 2015.

[20] Raminta Daniulaityte, Ramzi W Nahhas, Sanjaya Wijeratne, Robert G Carlson, Francois R Lamy, Silvia S Martins, Edward W Boyer, G Alan Smith, and Amit Sheth. time for dabs: Analyzing twitter data on marijuana concentrates across the us. *Drug and alcohol dependence*, 155:307–311, 2015.

[21] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.

[22] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *7th International AAAI Conference on Web and Social Media (ICWSM)*, 13:1–10, 2013.

[23] Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. Gender and cross-cultural differences in social media disclosures of mental illness. 2017.

[24] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

[25] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52. ACM, 1999.

[26] Virgile Landeiro Dos Reis and Aron Culotta. Using matched samples to estimate the effects of exercise on mental health from twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 182–188, 2015.

[27] Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, pages 20–24. Citeseer, 2013.

[28] Monireh Ebrahimi, Amir Hossein Yazdavar, and Amit Sheth. On the challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems*, 2017.

[29] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[30] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. Quote rts on twitter: usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, pages 200–204. ACM, 2016.

[31] Judith Gelernter and Wei Zhang. Cross-lingual geo-parsing for non-structured data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 64–71. ACM, 2013.

[32] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47. ACL, 2011.

[33] John F Gunn and David Lester. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, 17(3), 2015.

[34] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. Tracking suicide risk factors through twitter in the us. *Crisis*, 2014.

[35] Zongcheng Ji, Aixin Sun, Gao Cong, and Jialong Han. Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1271–1281. WWW, 2016.

[36] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2000.

[37] Christian Karmen, Robert C Hsiung, and Thomas Wetter. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer methods and programs in biomedicine*, 120(1):27–36, 2015.

[38] Revathy Krishnamurthy, Pavan Kapanipathi, Amit P Sheth, and Krishnaprasad Thirunarayan. Knowledge enabled approach to predict the location of twitter users. In *European Semantic Web Conference*, pages 187–201. Springer, Cham, 2015.

[39] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter data analytics*. Springer, 2014.

[40] Francois R Lamy, Raminta Daniulaityte, Amit Sheth, Ramzi W Nahhas, Silvia S Martins, Edward W Boyer, and Robert G Carlson. those edibles hit hard: Exploration of twitter data on cannabis edibles in the us. *Drug and alcohol dependence*, 164:64–70, 2016.

[41] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[42] Wendy Liu and Derek Ruths. Whats in a name? using first names as features for gender inference in twitter, 2013.

[43] Elsa Loekito and James Bailey. Using highly expressive contrast patterns for classification-is it worthwhile? *Advances in Knowledge Discovery and Data Mining*, pages 483–490, 2009.

[44] Shervin Malmasi and Mark Dras. Location mention detection in tweets and microblogs. In *International Conference of the Pacific Association for Computational Linguistics (PACL)*, pages 123–134. Springer, 2015.

[45] Riham Mansour, Mohamed Farouk Abdel Hady, Eman Hosam, Hani Amr, and Ahmed Ashour. Feature selection for twitter sentiment analysis: An experimental study. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 92–103. Springer, 2015.

[46] Stuart E Middleton, Lee Middleton, and Stefano Modafferi. Real-time crisis mapping of natural disasters using social media. *The Institute of Electrical and Electronics Engineers (IEEE) Intelligent Systems*, 29(2):9–17, 2014.

[47] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[48] Ahmed Mourad, Falk Scholer, and Mark Sanderson. *Language Influences on Tweeter Geolocation*, pages 331–342. Springer, Cham, 2017.

[49] Meena Nagarajan, Amit Sheth, and Selvam Velmurugan. Citizen sensor data mining, social media analytics and development centric web applications. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 289–290, New York, NY, USA, 2011. ACM.

[50] Meenakshi Nagarajan, Hemant Purohit, and Amit P Sheth. A qualitative examination of topical tweet and retweet practices. *4th International AAAI Conference on Web and Social Media (ICWSM)*, 2(010):295–298, 2010.

[51] Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. Proactive screening for depression through metaphorical and automatic text analysis. *Artificial intelligence in medicine*, 56(1):19–25, 2012.

[52] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226, 2014.

[53] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

[54] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *7th International AAAI Conference on Web and Social Media (ICWSM)*, 2013.

[55] Minsu Park, David W McDonald, and Meeyoung Cha. Perception differences between the depressed and non-depressed users in twitter. In *7th International AAAI Conference on Web and Social Media (ICWSM)*, volume 9, pages 217–226, 2013.

[56] Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. Retweet modeling using conditional random fields. In *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, pages 336–343. IEEE, 2011.

[57] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification, 2011.

[58] Sujan Perera, Pablo Mendes, Amit Sheth, Krishnaprasad Thirunarayan, Adarsh Alex, Christopher Heid, and Greg Mott. Implicit entity recognition in clinical documents. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 228–238, 2015.

[59] Sujan Perera, Pablo N Mendes, Adarsh Alex, Amit Sheth, and Krishnaprasad Thirunarayan. Implicit entity linking in tweets. In *Extended Semantic Web Conference (ESWC)*, pages 118–132, Greece, 2016.

[60] Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2015, page 21, 2015.

[61] Joseph D Prusa, Taghi M Khoshgoftaar, and David J Dittman. Impact of feature selection techniques for tweet sentiment classification. In *FLAIRS Conference*, pages 299–304, 2015.

[62] Hemant Purohit, Jitendra Ajmera, Sachindra Joshi, Ashish Verma, and Amit Sheth. Finding influential authors in brand-page communities. In *6th International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, 2012.

[63] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1), 2013.

[64] Hemant Purohit, Guozhu Dong, Valerie Shalin, Krishnaprasad Thirunarayan, and Amit Sheth. Intent classification of short-text on social media. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 222–228. IEEE, 2015.

[65] Hemant Purohit, Alex Dow, Omar Alonso, Lei Duan, and Kevin Haas. User taglines: Alternative presentations of expertise and interest in social media. In *2012 International Conference on Social Informatics (SocInfo), Washington, D.C., USA, December 14-16*, pages 236–243, 2012.

[66] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. pigeo: A python geotagging tool. *Proceedings of ACL-2016 System Demonstrations*, pages 127–132, 2016.

[67] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. ACL, 2011.

[68] Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. Twitter user gender inference using combined analysis of text and image processing. *V&L Net*, 2014:54, 2014.

[69] Roger C Schank. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631, 1972.

[70] Amit Sheth, Sujan Perera, Sanjaya Wijeratne, and Krishnaprasad Thirunarayan. Knowledge will propel machine understanding of content: Extrapolating from current examples. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 1–9. ACM, 2017.

[71] Amit Sheth, Hemant Purohit, Gary Alan Smith, Jeremy Brunn, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, and Wenbo Wang. Twitris: A system for collective social intelligence. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*, pages 1–23, New York, 05/2018 2018. Springer-Verlag New York.

[72] Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip S Yu, and Ming-Syan Chen. Mining online social data for detecting social network mental disorders. In *Proceedings of the 25th International Conference on World Wide Web*, pages 275–285, 2016.

[73] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter

network. In *2010 IEEE Second International Conference on Social computing (SocialCom)*, pages 177–184. IEEE, 2010.

[74] Paul Thompson, Chris Poulin, and Craig J Bryan. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–6, 2014.

[75] Laura Tolosi, Andrey Tagarev, and Georgi Georgiev. An analysis of event-agnostic features for rumour classification in twitter. In *10th International AAAI Conference on Web and Social Media*, 2016.

[76] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652. ACM, 2012.

[77] Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 90–94. ACL, 2012.

[78] Wenbo Wang, Lu Chen, Ming Tan, Shaojun Wang, and Amit P Sheth. Discovering fine-grained sentiment in suicide notes. *Biomedical informatics insights*, 5(Suppl 1):137, 2012.

[79] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE, 2012.

[80] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. Cursing in english on twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 415–425, New York, NY, USA, 2014. ACM.

[81] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213. Springer, 2013.

[82] Sanjaya Wijeratne, Lakshika Balasuriya, Derek Doran, and Amit Sheth. Word embeddings to enhance twitter gang member profile identification. In *IJCAI Workshop on Semantic Machine Learning (SML)*, pages 18–24, New York City, 07 2016.

[83] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. Emojinet: Building a machine readable sense inventory for emoji. In *8th International Conference on Social Informatics (SocInfo)*, pages 527–541, Bellevue, WA, USA, November 2016.

[84] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. Emojinet: An open service and api for emoji sense discovery. In *11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 437–446, Montreal, Canada, May 2017.

[85] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. A semantics-based measure of emoji similarity. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 646–653, 2017.

[86] Sanjaya Wijeratne, Derek Doran, Amit Sheth, and Jack L. Dustin. Analyzing the social media footprint of street gangs. In *IEEE International Conference on Intelligence and Security Informatics (ISI), 2015*, pages 91–96, May 2015.

[87] I. D. Wood and S. Ruder. Emoji as emotion tags for tweets. In *LREC 2016 Workshop on Emotion and Sentiment Analysis*, 2016.

[88] Ian. D. Wood and Sebastian Ruder. Emoji as emotion tags for tweeter. In *Proceedings of the LREC 2016 Workshop on Emotion and Sentiment Analysis*, 2016.

[89] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM, 2011.

[90] Amir Hossein Yazdavar, Hussein S Al-Olimat, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Amit P Sheth. Analyzing clinical depressive symptoms in twitter. 2016.

[91] Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Sydney, Australia, 2017.

[92] Shaozhi Ye and Shyhtsun Felix Wu. Measuring message propagation and social influence on twitter. com. *2nd International Conference on Social Informatics (SocInfo)*, 10:216–231, 2010.