A Survey on Using Social Media as a Sensor
PhD Qualification Exam Report, Summer 2015

Zahra Iman (imanz@onid.oregonstate.com)
Oregon State University

Abstract

Contents

1 Introduction

Sensors are devices used for measuring some aspect of an environment and converting it into continuous or discrete value for use in information or control systems. For example, thermocouple as a sensor senses the temperature and converts it to continuous output voltage or a fire detector outputs a boolean value as a result of sensing smoke. Expanding on this definition, social media can be used as a sensor to detect important news e.g., death of Michael Jackson, real-time events e.g., earthquake or epidemics, sentiments and opinions e.g., people's political alignment towards political debates, preferences and traits e.g., stock market prediction or purchase behavior prediction.

However, due to the complexity of information in social media, designing sensors to detect specific phenomena requires highly specialized research to extract targeted information. Designing methodologies and extracting features that are robust to highly dynamic changes in social media, with various forms of expression e.g., informal, short, unstructured texts, written by individuals from different educational levels, and with large volumes of extraneous material mixed in is a difficult task in need of extensive research.

This paper surveys existing work to explore social media as a sensor, methodology and features developed. This paper concludes by examining open areas for future research.

2 Use Cases of Social Media Sensor

This section provides existing research on the use of social media as a sensor for three major use cases of events, sentiment, and preference. This information is important for different target users such as government agencies, traffic incident management departments, marketing companies, and individuals.

2.1 Events

A social media event can be defined as an occurrence at a certain time interval and geographical region. It can be planned or unexpected e.g, concert vs. death of a celebrity, man-made or natural e.g., parade vs. earthquake, local or global e.g., concert vs. World Peace Day. Events can further be categorized based on their target users, including individuals, government agencies concerned about natural disasters and health epidemics, marketing companies, and news websites.

Historically, event detection has been studied extensively in text mining, NLP, and IR to find events from conventional media sources such as news streams Yang et al. [70]. With the growth of social media sites such as Facebook, Twitter and other microblogs, social media sites have become known as powerful communication tools for sharing and exchanging information about such events. However, event detection on social media sites is more challenging due to features such as unstructured and informal text, highly length restricted, generated by novice reporters compared to journalism-trained news editors.

Nevertheless, it is important to investigate event detection in social media because in comparison to traditional news blogs, social media has faster response time to events and time is money (marketing), lives (disasters), or simply relevance (new).

To see how different use cases address the aforementioned technical difficulties, we focus on the three highly studied types of event detections:

- Trending Topic Detection
- Natural Disaster Detection
- Health Epidemic Detection

In the next section, we summarize the results of trending topic detection research.

2.1.1 Trending Topic Detection

Trends i.e. emerging topics are typically driven by emerging events, breaking news and general topics such as death of celebrities, festivals, and sporting events that attract the attention of a large fraction of Twitter users [40]. Real-time detection of events which are hypothesized to be trendy is thus of high value for news reporters and analysts.

The following works on detecting trending topics use bursts as the indicator of events, where a burst is defined as a sudden change in posting rates of some keywords, hashtags, etc. However, they can be divided into multiple categories based on how they use bursts to extract the event.

Clustering-based Methods This category of works focus on the hypothesis that trends are topical and topics are defined by collection of relevant content, hence trends can be detected by clustering posts.

- **Threads of tweets** Petrovic et al. [50] tried to detect novel events from streams of Twitter posts by forming threads of similar tweets. The minimum similarity distance to an existing tweet represents novelty score of the tweet. Similarity threshold for assigning tweets to threads controls size of threads. The fastest growing thread in each time interval indicated the news of the event spreading and are outputted as new events.
- **Wavelet analysis** Weng and Lee [67] applied wavelet analysis to individual words on the frequency based raw signals of the words and identified events by grouping a set of words with similar burst patterns.

Term-based Methods The second category of works focus on the hypothesis that topics can be detected by focusing on temporal patterns of terms/keywords independent of contents of documents.

- **Keyword-burst** Mathioudakis and Koudas [40] detected events by focusing on bursts of keywords whereas Cui et al. [17] used different hashtag properties to this purpose. Zhao et al. [74] and Nichols et al. [44] also tried to use bursts in keywords, but they monitored specific keywords related to sports game in order to detect important NFL games or important moments within the game.

Term-based Methods The third category of works focus on detecting trending topics by measuring trendiness factor based on user-defined criteria.

- **Location-dependent** Emphasizing location, Albakour et al. [2] and Sakaki et al. [55] detected local events based on either user query, user location, or both. Albakour et al. employed contents of the tweets and volume of microblogging activite for locating events in a local area and ranked tweets on the level of topical relevancy to user query resulting in ranked list of local events. Sakaki et al. used classification approach for detecting driving events at a local area by using dependency of words to search query, context (words before or after a search query), position of a search query in a tweet, time expression in a tweet, and word features (all words in the tweet) as features.

- **Structure-based** Budak et al. [12] incorporated network topology in order to find trending topics. They defined trendiness of a topic based on two notions, either by the number of connected pairs of users discussing it, or by scoring a topic based on the number of unrelated people interested in it.

2.1.2 Natural Disaster Detection

In case of disasters, users will tweet about the disaster within seconds of its happening [1]. Using this information, disasters can be detected almost in real time from social media and responded to by government agencies such as U.S. FEMA (Federal Emergency Management Agency), local first responders, news websites, and individuals. The goal of works targeting disastrous events on Twitter can be divided into the two following categories:

Predictive studies on disaster Kryvasheyeu et al. [37] studied the network of users and focused on choosing the best groups of users in order to achieve lead-times i.e. faster detection of disastrous event (following the concept of "friendship paradox"[2]). On the other hand, Sakaki et al. [56] used SVM classifier for detecting earthquake and employed location estimation method such as Kalman Filtering for localizing it. Sakaki et al. extracted statistical features e.g., the number and position of words in a tweet, keyword features and word context features. These studies investigated the real-time nature of Twitter and provided promising results.

Descriptive studies on disaster Related works discuss the behavior of Twitter users

[1] http://mashable.com/2009/08/12/japan-earthquake/

[2] On average, most people have fewer friends than their friends have

during crisis [16, 60, 63] but do not address exploiting detection of crisis events. They investigated the use of social media during crisis in order to identify information propagation properties, social behavior of users e.g. retweeting behavior, information contributing to situational awareness, and active players in communicating information. However, this behavioral information could be exploited in development of sensors.

2.1.3 Health Epidemic Detection

A disease outbreak can rapidly infect great numbers of people and expand to broad areas involving several countries such as Ebola [3]. It is very important to identify the infected sources as early as possible and control the spread of epidemics by incubating infected individuals [15, 21]. Target users of this event detection include government agencies such as the CDC (Centers for Disease Control and Prevention), news websites, and individuals.

The purpose of these works was early detection of outbreaks using tweets. Researchers used content-based method and/or structure-based methods outlined as follows:

Content-based methods Culotta [18] and Aramaki et al. [3] both tried to identify influenza-related tweets and find correlations of these tweets to CDC statistics. Both works extracted bag-of-words as features. As for methodology, the former used single and multiple linear regression showing that multiple linear regression works better, while the latter employed SVM. Results showed high correlation of their estimation of influenza in early stages with values from U.S CDC and Japan's Infection Disease Surveillance Center.

Structure-based method In contrast to previous methods, García-Herranz et al. [24] focused on a structure-related technique and developed a model for contagious information diffusion in a social network. They provided a method for choosing sensor groups from friends of random sets of users to find more central individuals in order to enforce early detection.

Hybrid method Sadilek et al. [54] exploited both content and structure information. They employed a semi-supervised approach to learn a robust SVM classifier by training two classifiers on a labeled set and then applying them to non-labeled tweets . This enabled them to model the interplay of social activity, human mobility, and the spread of infectious disease in a large real-world population. Further, **?** identified sick individuals from the content of online communication.

2.2 Sentiments and Opinions

Sentiment analysis, also known as opinion mining, is defined as analysis of text based on expressed sentiments by users. Users share their opinions about products, political

[3] http://www.cdc.gov/vhf/ebola/outbreaks/index.html

matters, the stock market, and pharmaceuticals [4]. Marketing companies, government agencies, and individuals are concerned with what users think about them/their products. The goal here is to learn the model of users' sentiment toward these matters. To this purpose, different classification and statistical methods are used that are mentioned in the following sections.

2.2.1 Types of Sentiment Analysis

Two major aspects of sentiment analysis are the following:

Subjective vs. Objective sentiment At the top level of analysis, sentiment can be classified as subjective or objective [38]. Subjective text indicates a writer's opinion or emotional state with respect to some topic e.g., "it's an excellent phone", while objective text indicates a desirable or undesirable condition e.g., "it is broken".

Simple vs. Complex sentiment Simple sentiment shows whether a text's attitude is positive or negative. Complex sentiment involves the sentimental reaction of the human to various words across different factors, such as measuring the scale of positivity/negativity, potency, oriented activity, receptivity, aggressiveness, novelty, and tension and will be discussed in the applications of sentiment analysis.

Regardless of the features and sentiment type, sentiment analysis in social media has different application which are discussed in more detail in the following sections.

2.2.2 Framework of Sentiment Analysis

Methods used in sentiment analysis included:

- Classification e.g., SVM, NB, Maximum Entropy, Neural Network
- Manual Coding and Concept Extraction
- Simple Statistical Methods
- Textual Analysis Softwares

$$Methods \begin{cases} Classification\ e.g.\ SVM,\ NB,\ Maximum\ Entropy,\ Neural\ Network \\ Manual\ Coding\ and\ Concept\ Extraction \\ Simple\ Statistical\ Methods \\ Textual\ Analysis\ Softwares \end{cases}$$

[4] pharmacological science relating to the collection, detection, assessment, monitoring, and prevention of adverse effects of pharmaceutical products

$$Features \begin{cases} n-grams\ e.g.,\ uni-gram,\ bigram,\ and\ trigram \\ POS-Tags \\ Negations \\ Colocation\ and\ Position\ of\ words \\ Semantic\ features\ e.g.,\ extracting\ concept\ from\ medical\ dictionary \end{cases}$$

2.2.3 Applications of Sentiment Analysis

Political Applications Social media has been extensively used during political events. For example, analysts attribute Obama's victory to the strength of his social-networking strategy and use of social media such as mybarackobama.com, or MyBO [62] which shows the extent and influence social networking holds during political debates and events.

Researchers have studied social media in order to either investigate and evaluate the relationship of online political sentiment to offline political landscape [5, 45, 62, 65] or to see if online political sentiment can be predictive of actual election results [6, 41]. Methodologies used for these purposes include using textual analysis software (LIWC [49]) [62], classification e.g., Naive Bayes, SVM, Adaboost) [5, 6, 41, 65], or simple statistical methods such as computing sentiment score as the ratio of positive to negative word counts [45]. These methods are based on different sets of features extracted from text such as lexicon-based features [5], the frequency of keywords [45, 62], and with uni-grams being the most commonly used feature [6, 41, 65].

Regarding the predictive power of Twitter, [6, 41] extracted simple sentiment from social media and compared it to actual national polls results. Bermingham and Smeaton [6] claim that social analytics using both volume-based measures and sentiment analysis are predictive of public opinion during the Irish general election. On the other hand, Mejova et al. [41] argue that online sentiment is not predictive of national poll results on US presidential candidates. Tumasjan et al. [62] went further and extracted complex sentiment for 12 emotional dimensions for profiling political sentiment about parties in the parliament. They showed that the mere number of messages mentioning a party reflects the election result.

Product Market Applications Everyday, social media users comment and share their opinions about different products. Extracting useful information from these opinions are helpful to marketing companies, news websites, and individuals.

Research in this application area targets different products, e.g., movies, laptops, cameras, books, music [19, 48], trends of different brands in social media, and the relationship between the company and customers [25, 32]. Current research takes advantage of off-the-shelf classifiers e.g., SVM, Naive Bayes, Maximum Entropy, and Neural Networks in order to classify product reviews into simple sentiment i.e. positive, negative, or neutral. Different features have been extracted to this purpose. While all of these works share uni-grams as features, Pang et al. [48] used POS-tags and position of words, Dave et al. [19] used other linguistic features e.g., negations and colocation, and Ghiassi et al. [25]

extracted emoticons in addition to n-grams.

Moreover, in contrast to [19, 48] who extract simple sentiment, [25, 32] use graded sentiment on a 1 to 5 scale to rank sentiment toward brands. Results suggest that people do tweet about different brands and products and these works were able to extract the sentiments about them with reasonable accuracies. They compared the classification results to scalar rating per product provided in the websites such as Amazon, IMDB, etc.

Stock Market Applications Bollen et al. [10] took advantage of Google-Profile of Mood States (GPOMS) to extract 7 public mood time series, in addition to simple positive, negative sentiment, to see if public mood is predictive future of stock market values. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network trained on the basis of past DJIA values and public mood time series were used to investigate the hypothesis that public mood states are predictive of changes in stock market closing values. The econometric technique of Granger causality analysis is applied to the daily time series produced by GPOMS vs. the DJIA[5]. Granger causality analysis rests on the assumption that if a variable X causes Y then changes in X will systematically occur before changes in Y. Each public mood time series is then compared to DJIA time series with these methods to observe the predictive power of the mood. Specifically, they claimed that the calmness of the public (measured by GPOMS) was predictive of stock market values. Inline with this finding, Zhang et al. [71] also showed that Twitter posts can be used to predict market indices.

Pharmacovigilance Applications Another application of sentiment analysis on social media belongs to the study of online posts for monitoring of Adverse Drug Reactions (ADR). ADR research has focused on social media due to its large volume of user-posted information.

Researchers have investigated Twitter posts looking for potential signs of ADR [33, 47] and/or to identify potential users of drugs [7]. Methodology used in these works is similar to product market research and includes typical classification methods e.g., SVM and Maximum Entropy [7, 33], and manually coded classification with concept extraction and lexicon matching [47] in order to detect mentioned signs of ADR in posts. These methods are based on various features extracted from posts such as semantic features generated by MetaMap[6] concerning mention of ADRs [7, 33, 47], presence and frequency of semantic types of disease or syndrome [33], and textual features e.g., number of hashtags, reply-tags, urls, pronouns [7, 33]. Results suggest that users mention adverse drug reactions and studying social media data can serve to complement and/or supplement traditional time-consuming and costly surveillance methods [33].

[5] A price-weighted average of 30 significant stocks traded on the New York Stock Exchange and the Nasdaq

[6] A program mapping biomedical text to concepts in the largest thesaurus in the biomedical domain [4]

2.3 Preferences and Traits

There are two types of preference learning problems. The first is where there is only a single user and many items. Usually, researchers use product description as features of the items in order to do the prediction and the predictions are based on ranking items. The second case is when there are multiple users and multiple items. This scenario is often called collaborative filtering [31]. What do users prefer to buy, who do they prefer to be the next president, what pages would they like, what topics are the most interesting ones for them, and what are their private traits? All of these questions can be answered through learning the user's preferences. Marketing companies and political parties are the most important target users of this procedure .

2.3.1 Framework of Preference Prediction

Predicted preferences can be absolute or relative. Absolute preferences are further divided into binary or numeric e.g. U_1 rates X_2 as 3 or $Rating(U_1, X_2) = 3$. Relative preferences show ordering on a set of items e.g. $X_1 \succeq X_2 \succeq X_3$. Four different methodologies are commonly used:

- Content-based: methods based on features extracted from the content of posts by employing simple linear regression, classification, or data mining approaches
- Interaction-based methods: depend on the links and interaction between users (share, comment, tag, mention, like, retweet)
- Collaborative Filtering: methods aiming to exploit information about preferences for items, including matrix factorization and neighborhood models
- Hybrid: any of the above methods using content and interaction information to extract preferences by employing simple linear regression, classification, or data mining approaches

2.3.2 Applications of Preference and Trait Prediction

This section provides various works on preference learning and trait prediction.

Traits and Personal information prediction Studies in this section provide predictions for users' personality traits, intelligence, gender, age, sexual orientation[36] or extract characteristics of users. For example, they show that there is a correlation between popularity (measured by following, followers, and listed counts on Twitter profile) and extroversion (measured by myPersonality test[7]) shown with computation of Pearsons correlation[53]. Method used by [36] and [53] are both interaction-based. Numeric variables such as age or intelligence were predicted using a linear regression model, whereas

[7] http://www.mypersonality.org/wiki/

dichotomous variables such as gender or sexual orientation were predicted using logistic regression. Kosinski et al. [36] used Facebook likes as the only feature, while Quercia et al. [53] used more extensive features including user's profile information, number of followers, and number of followees.

Product preference prediction/ Product recommendation Research on product preference prediction targets different products such as electronics, movies, music, and foods. Researchers provided various types of output including a ranked list of products [72, 73], numeric real-values showing the preferences for each item [58], or binary values on whether the user would like an item or not [57]. They used different methodologies such as simple popularity methods [72], linear regression [57, 72, 73], simple classifiers (Naive Bayes, SVM, logistic regression) [57, 72], or collaborative filtering methods based on matrix factorization [58]. Zhang and Pennacchiotti [72, 73] uses a set of features derived from the users social media account, e.g., Facebook page likes and user demographics, Facebook n-grams from pages and user's purchase behaviors from e-bay. Sedhain et al. [57] focuses on user interactions (type, modality, directionality) in addition to user likes on Facebook.

Political preference prediction Research on political preferences includes predicting political orientation [26, 27], classifying stances on political debates concerning topics of health care, gay rights, gun rights, ... [59, 64], or providing descriptive study on users' influences on political orientation of others [1] . Methodologies used are divided into collaborative filtering methods and non-collaborative methods.

- Gottipati et al. [27] applies collaborative filtering based on probabilistic matrix factorization.
- The non-collaborative works either use simple data mining and statistical approaches [26], homophily measure between users and their followers/followees using similarity metrics [1], or classification methods [59, 64]. To apply these methods, researchers extracted features including sentiment features [59, 64], and structure-based features (network of users on following each other) [1, 26].

Results suggested (with highest accuracy of 70%) that it is possible to detect the stance of users toward political debates or parties. The descriptive study of [1] showed that in 73% of cases, users and their followers shared similar political orientation.

2.3.3 Re-Tweet Prediction

Information diffuses in Twitter between users through retweets. Analyzing retweet history reveals users personal preference for tweets. Therefore, predicting retweet behavior of a tweet and studying characteristics of popular messages are important for understanding and controlling information diffusion in Twitter.

To this end, various works have been proposed with three different main goals:

1. **Predict if a tweet will be retweeted in future and provide retweet count** [14, 51, 69]: All of these works use classification-based approaches using tweet-based and author-based features. However, Can et al. [14] took advantage of visual cues from images linked in the tweets, and Xu and Yang [69] employed social-based features in addition to tweet and author-based features. In contrast to these methods focused on the entire dataset, Petrovic et al. [51] worked on retweet prediction of real-time tweeting and employed online learning algorithms.

2. **Rank tweets based on retweeting probability or category** [23, 29]: Unlike the authors in part 1, Feng and Wang [23] used a general graph model by building a graph of users, publishers, and tweets as nodes represented with feature vectors. A feature-aware factorization model was proposed to re-rank the tweets by unifying the linear discriminative model and low-rank factorization model. Moreover, [23] used author-based and interaction-based features in addition to tweet-based features, while Hong et al. [29] used topological features and tweet-based features.

3. **Study important factors regarding retweet behaviors** Hong et al. [29] and Petrovic et al. [51] provided factors influencing information diffusion. Specifically, Hong et al. claimed that *degree distribution* and *retweet before* contribute greatly to retweet behavior. Petrovic et al. built separate, time sensitive models based on tweet's creation time. They showed that this substantially improved retweet predictions.

Most of these works used classification techniques e.g., SVM, linear regression, logistic regression, and random forest. The features used by these works included tweet-based, author-based, social-based, interaction-based, visual cues, and topological features. These features are described in more detail in table 1.

Feature Type	**Detail Features**
Tweet-based	TF-IDF, topics extracted from LDA, #urls, #hashtags, #users_mentioned, type (reply/retweet), #total_words, has_multimedia, has_geography, time-span since last rt, time-span since created, tweet_length
Author-based	#followers, #friends, #tweets_published_before, #listed_times, #favorited_times, age, avg #tweets per day, location, is_verified
Social-based	Author relationship to user: is_followed, is_in_list, #times_retweeted, is_followee, #times_mentioned
Interaction-based	tweet profiles similarity, recent tweet profiles similarity, reply_count, self-descriptions similarity, following lists similarity, retweet_count, has_same_location/timezone, mention_count,
Visual cues	GIST, color histograms
Topological	Page-rank, degree distribution, local clustering coefficient, reciprocal links

Table 1: List of features used in retweet prediction

3 Supporting Theories

We began by discussing existing research on applications of using social media as a sensor, now we discuss a range of theories that support these applications.

3.1 Sentiment Theories

Sentiment theories cover the characteristics of text necessary for determining the attitude of the author. Attitude can be based on author's judgment [68], affective or emotional state [46], or the intended emotion the author wanted to give [30]. The first one gets the emotion from author's point of view on a subject, second one conveys the state of the author at the time of writing, and the last one is the emotional effect that the author was trying to convey to the reader. This is the reason that research should include these complications into sentiment analysis. For example, using humor in regards to product review e.g., this is great, it could not be any better, It broke in two days.

Here, we provide an overview of complex and simple sentiment theory, appraisal theory, and linguistic theories on how people write about their emotions. These theories empower sentiment analysis tools to extract the emotions from text fro various applications outlined in section 2.2.

3.1.1 Complex and Simple Sentiment

As was mentioned earlier, sentiment analysis can be simple and analyze polarity of text as being positive or negative, or be complex and extract multi-dimensional sentiments.

There are a few different major theories of complex sentiment [13], outlined as follows:

Sentimental Reaction to Various Words Osgood [46], in a study of text polarity showed human's sentimental reaction to various words across eight dimensions:

- Evaluation (positive or negative)
- Potency (strong or weak)
- Activity (active or passive)

Each aspect is characterized by a variety of contrasts. Characterizations of the positive side of each dimension is [28]:

- Evaluation: nice, sweet, heavenly, good, mild, happy, fine, clean
- Potency: big, powerful, deep, strong, high, long, full, many
- Activity: fast, noisy, young, alive, known, burning, active, light

The corresponding words for the negative side are

- Evaluation: awful, sour, hellish, bad, harsh, sad, course, dirty
- Potency: little, powerless, shallow, weak, low, short, empty, few
- Activity: slow, quiet, old, dead, unknown, freezing, inactive, dark

Appraisal Theory Appraisal theory is the psychological theory arguing that emotions come from our subjective evaluation and interpretation (appraisals or estimates) of events. Each appraisal expression has three main components: an attitude (which takes an evaluative stance about an object), a target (the object of the stance), and a source (the person taking the stance) which may be implied [68]. In general, appraisal theory is an analysis of how a writer values people and things within the text that he/she produces [39]. It studies different types evaluative language that can occur and represents three grammatical systems comprising appraisal [8]:

- Attitude: tools that an author uses to directly express his approval or disapproval of something and it is divided further into:
  - affect (internal emotional evaluation of things)
  - judgment (evaluation of a person's behavior within a social context)
  - appreciation (aesthetic or functional evaluation of things)
- Engagement: resources which an author uses to position his statements relative to other possible statements on the same subject such as claims, states, informs, ...
- Graduation: resources which an author uses to convey the strength of that approval or disapproval such as very, reasonably, ...

Hence, this theory and Osgood [46]'s theory are parallel to each other (attitude/evaluation, potency/graduation) on some aspect and differ from each other on the other aspects (activity, engagement). These theories have been used in different sentiment analysis works such as [34, 43] for classifying words.

Psycho-Linguistic Theories The third theory focuses more on psychological aspect of language and how people with different psychological background use words. It differs from the last two theories in the way that it is more general and focuses specifically on the usage of different types of words in different positions in the sentence and how they relate to different emotional indicators. Tausczik and Pennebaker [61] show linguistic theories and psychological evidence behind them. They reviewed several text analysis methods to support the hypothesis that people provide enough clues in their language to enable us to detect their feeling and emotions. It is possible to relate daily word use to a broad array of real-world behaviors shown in table **??**. Various word usages include:

- nouns
- adjectives

- adverbs
- style words (pronouns, prepositions, articles, conjunctions, auxiliary verbs)
- verb tenses
- positive/negative emotion words
- word count
- point-of-view (first, second or third person)
- exclusion words
- causal words

Different emotional indicators based on the word usages include:

- emotional state
- social relationships
- thinking style
- individual differences
- social hierarchy
- social coordination
- honesty or deception
- attentional focus

3.2 Social Network Theories

Section 2 covered the existing literature on detection of trending topics, disasters, health epidemics, preferences, and traits. However, what are the characteristics of social networks that enable diffusion of information, ideas, or what features do neighbors hold in the network? For example, awareness as the most important aspect of information flow during a disaster is related to the centrality concept in social networks. Sentiment is one useful theory for various targets. Social network analysis provides another additional theory and explores architecture of social networks and explores phenomena that are generated according to structure, information flow.

3.2.1 Graph Structure

A major subset of Social Network Analysis is graph structure analysis. This section provides studies on how social network graphs are generated, how information flows through social networks, and how different users play structurally distinct roles. Moreover, we show the importance of certain topological properties of networks, such as the concept of weak ties, the number of social connections that an individual has in a given society, or the number of communities that a society forms [20, 66]. We first provide some basic properties in networks and then we discuss different graph generation models that provide generative models of social network graph that reproduce these properties.

Basic Concepts

- **Clustering coefficient** Measures the probability that two randomly chosen friends of a user are friends themselves.
- **Strong and Weak ties** Weak ties are links in the network that connect two users with no common friend which enables linking different tightly-knit communities. In contrast to this, links between these tightly-knit communities represent strong ties. The importance of the weak ties lies in the fact that it can provide the involving users with access to different parts of network that otherwise would have been inaccessible. Figure 1 shows this concept.
- **Triadic Closure** The property among three nodes A, B, and C, such that if a strong tie exists between $A-B$ and $A-C$, there is a weak or strong tie between $B-C$.
- **Centrality** Characterizes the importance of nodes (individuals). The degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, Katz centrality (PageRank), precolocation centrality, and cross-clique centrality of a node are related measures regarding the importance of nodes.

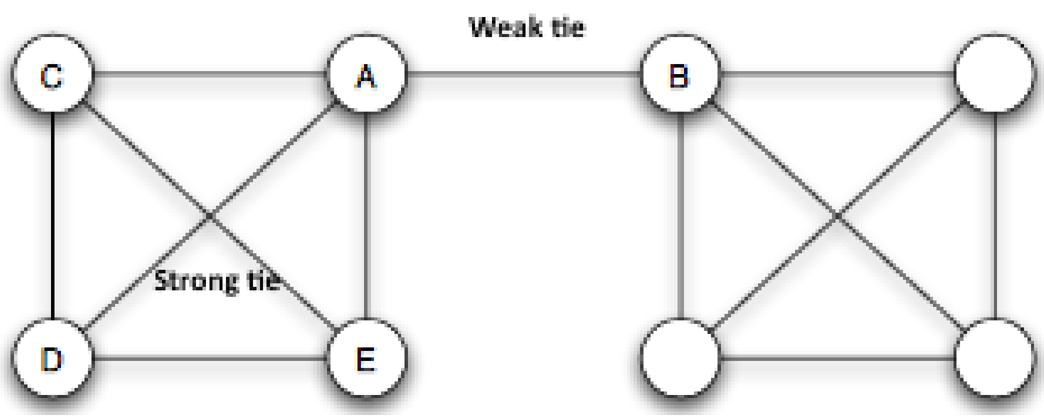

Figure 1: Weak vs. Strong ties: edge A-B is a weak tie while edge C-E is a strong tie

Graph Generation Models Historically, three different graph models have been studied:

1. Random graphs: graphs generated by starting with a disconnected set of nodes that are then paired with a uniform probability

2. Watts and Strogatz graph: graphs with small-world properties[8], including short average path lengths and high clustering

3. Scale-free networks: graphs characterized by a highly heterogeneous degree distribution, which follows a "power-law"

4. The Barabasi-Albert model: the first network with a power-law distribution

These graphs are represented as $G = (V, E)$, with V showing the set of Vertices e.g, people and E corresponding to edges e.g., friendship relationship. A path consists of a set of edges connecting two nodes together. There are three important concepts regarding reproduction of complex social network structure:

1. Average path length: showing the average value of length of different paths that characterizes a network's compactness.

2. Degree distribution: the probability distribution of degrees over the network

3. Clustering coefficient

Random networks, also known as Erdos-Renyi networks, are an entirely random network based on a probability p of connecting nodes. They have short path length and independent edges. Social networks are not random graphs since they represent preferential attachment and small world behavior. Preferential attachment refers to the observation that in networks that grow over time, the probability that an edge is added to a node with d neighbors is proportional to d. The concept of small-world networks introduced by Watts and Strogatz (WS) characterizes networks in which most nodes can be reached from every other in a small number of steps (following the six degree of separation theory).

Unlike the last two static structures, scale-free networks are dynamically formed by continuous addition of new nodes to the network, and don't have uniform probabilities for creation of new edges. The two main ingredients for self-organization of a network in a scale-free structure are growth and preferential attachment. The corresponding degree distribution is a power law. These networks have smaller average path length compared to random graphs and small-world networks [66]. **?**] introduced an algorithm for generating a network with power-law distribution and having two ingredients of growth and preferential attachment. Growth is the concept regarding the observation that most real-world networks describe open systems that grow by the continuous addition of new nodes. An important corollary of graph structure is shown below.

Friendship Paradox The important concept of friendship paradox explained below is derived from graph generation models and their properties. Feld [22] introduced the concept of friendship paradox. Using general mathematical properties of social networks, he showed that on average most people have fewer friends than their friends have.This phenomenon was explained as a consequence of the general mathematical properties of social networks. Assuming the graph of social network $G = (V, E)$ with V showing the

[8]most nodes can be reached from every other node by a small number of steps

set of people and E corresponding to friendship relationship. He modeled the average number of friends of a person in the social network as the average of the degrees of the vertices in the graph. And the average number of friends that a typical friend has, was modeled by choosing uniformly at random an edge and an endpoint of that edge, followed by calculating the degree of the selected endpoint again. Hence, Feld proved that your friends have more friends than you do.

3.2.2 Information Diffusion and Cascades

In this section, we provide studies on how social network structures support the diffusion of information. It will be shown that the topology of a network has great influence on the overall behavior of information cascades and epidemic spread.

Diffusion Models Given a social network and estimates of reciprocal influence, viral marketing aka influence maximization problem is defined to target the most influential users in the network in order to activate chain-reaction of influence and eventually influence largest number of users in the network. Two basic classes of diffusion models exist. These models represent social network as a directed graph with nodes (persons) starting as either active or inactive. Each active node may trigger activation of neighboring nodes.

1. Linear Threshold Model: each node has random threshold $\theta_v \sim U\left[0,1\right]$, and is influenced by each neighbor according to some weight. It becomes active if θ_v fraction of its neighbors are active.
2. Independent Cascade Model: if a node becomes active, it has a single chance of activating each currently inactive neighbor. The activation attempt succeeds with a certain probability related to those two nodes.

Epidemic spread In classic epidemiology individuals have an equal chance of contact. However, this was determined to be unrealistic. In response,Brede [11] introduced an underlying contact networks model [11]. Contact network represents each person as a node and contact as edges and the network can change based on the pathogen. Probability of contagion and length of infection is controlled by the contact network structure. A node will become infected if and only if there is a path to the node from one of the initially infected nodes. [42]

Famous epidemiological models have been introduced on contact networks:

1. SI: Susceptible-Infected [11]
2. SIR: Susceptible-Infected-Removed [35]
3. SIS: Susceptible-Infected-Susceptible

These models show potential stages individuals would go through and they model number of individuals in each stage as random variables. Infection rate β is defined as the probability of contagion after contact per unit of time. In general, patterns of epidemic spread depend on a disease's contagiousness.

In homogeneous networks where individuals have k possibilities of getting the contagion from neighbors, growth of epidemics depends on β and k and therefore depends on an epidemic threshold. Different from that, scale-free networks have a different equation for all nodes of same degree k, instead of assuming homogeneous mixing. Nodes with higher degrees are more susceptible to infection. Finally, we can observe that there is no epidemic threshold for (infinite) scale-free networks. In practice, the epidemic threshold in scale-free networks is going to be very small and finite-sized scale-free networks are susceptible to epidemic spread regardless of spreading rate.

Similar to this, social contagion phenomena refer to various processes that depend on the individual propensity to adopt and diffuse knowledge, ideas, information. Similar to epidemiological models:

- Susceptible: an individual who has not learned new information
- Infected: the spreader of the information
- Recovered: aware of information, but no longer spreading it

Two main questions are: 1) if the rumor reaches high number of individuals i.e. go viral, 2) rate of infection spread

4 Conclusions and Discussion of Open Areas

In conclusion, through review of the literature we showed that social media can be used as a sensor to detect latent phenomena. These works successfully detected or predicted events, sentiments, and preferences of users. The analysis of various works for each use case contributed information on prominent trends in analysis of social media. They started by extracting substantial sets of features and applied different statistical, data mining, natural language processing, or classification methods.

Results showed that different methods can be successfully applied for each task. For example, classification, regression, and collaborative filtering methods each could provide satisfactory results for some tasks of sentiment analysis, event detection, and preferences.

However, it was shown that the results of some tasks such as detection of political preferences provide low accuracies. In general, there is still a gap in detection of these phenomena from social media. The survey results indicate that there is a lack of smart methods for selecting user and learning better sensors compared to existing basic and off-the-shelf classifiers.

Therefore, future expansions of this area includes:

- Learning tunable sensors for each specific task such as detection of events, trending topics, disasters, epidemic, sentiment, and preference. A large body of work in detection of these tasks define use of sensors, but rely on ad-hoc methods based on data mining or simple statistics to select sensors. In contrast, we intend to learn tunable sensors for each task.
- Investigate the inclusion of temporal relevance in sentiment analysis methods. People tend to change their emotion and feeling toward topics over time. However, no one has yet to model the dynamics of opinion. We will focus on considering temporal dimension of data in our work.
- Co-training/learning error rates of independent sensors on large unlabeled data [9, 52]. This would enable us to take advantage of semi-supervised learning and, in turn, use all data including unlabeled data.

Finally, we establish a current state of practice in use of social media as a sensor. We discussed the theories supporting this concept and their abilities to answer the questions of how we can detect latent phenomena in real-time and what features of social networks and text allow us to do so. We surveyed use cases covering the use of social media to detect trending topics, natural disasters, health epidemics, sentiments, preferences, and individual traits. What remains is the need for thorough research into learning tunable sensors, the dynamics of user's opinion, and the learning error rates of independent sensors.

References

[1] Mohammad Ali Abbasi, Reza Zafarani, Jiliang Tang, and Huan Liu. Am i more similar to my followers or followees?: analyzing homophily effect in directed social networks. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*, pages 200–205, 2014.

[2] M-Dyaa Albakour, Craig Macdonald, and Iadh Ounis. Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 173–180, Paris, France, France, 2013. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. ISBN 978-2-905450-09-8.

[3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.

[4] Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In *AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001*, 2001.

[5] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics, 2013.

[6] Adam Bermingham and Alan F Smeaton. On using twitter to monitor political sentiment and predict election results. *In Proceeding of IJCNLP conference, Chiang Mai, Thailand*, 2011.

[7] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, SHB 2012, October 29, 2012, Maui, HI, USA*, pages 25–32, 2012.

[8] Kenneth Bloom. *Sentiment analysis based on appraisal theory and functional local grammars*. PhD thesis, Illinois Institute of Technology, 2011.

[9] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[10] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *J. Comput. Science*, 2(1):1–8, 2011.

[11] Markus Brede. *Networks - An Introduction*. mark e. j. newman. (2010, oxford university press.) \$65.38, £35.96 (hardcover), 772 pages. ISBN-978-0-19-920665-0. *Artificial Life*, 18(2):241–242, 2012.

[12] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Structural trend analysis for online social networks. *PVLDB*, 4(10):646–656, 2011.

[13] Sarah Bull. Automatic parody detection in sentiment analysis. 2010.

[14] Ethem F. Can, Hüseyin Oktay, and R. Manmatha. Predicting retweet count using visual cues. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1481–1484, 2013.

[15] CDC. Principles of epidemiology, second edition. *Centers for Disease Control and Prevention*.

[16] France Cheong and Christopher Cheong. Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In *Pacific Asia Conference on Information Systems, PACIS 2011: Quality Research in Pacific Asia, Brisbane, Queensland, Australia, 7-11 July 2011*, page 46, 2011.

[17] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover breaking events with popular hashtags in twitter. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 1794–1798, 2012.

[18] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0217-3.

[19] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3.

[20] David A. Easley and Jon M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[21] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[22] Scott L Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477, 1991.

[23] Wei Feng and Jianyong Wang. Retweet or not?: personalized tweet re-ranking. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 577–586, 2013.

[24] Manuel García-Herranz, Esteban Moro Egido, Manuel Cebrián, Nicholas A. Christakis, and James H. Fowler. Using friends as sensors to detect global-scale contagious outbreaks. *PloS one*, abs/1211.6512, 2012.

[25] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Syst. Appl.*, 40(16):6266–6282, 2013.

[26] Jennifer Golbeck and Derek L. Hansen. A method for computing political preference among twitter followers. *Social Networks*, 36:177–184, 2014.

[27] Swapna Gottipati, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. Predicting user's political party using ideological stances. In *Social Informatics - 5th International Conference, SocInfo 2013, Kyoto, Japan, November 25-27, 2013, Proceedings*, pages 177–191, 2013.

[28] David R Heise. *Expressive order: Confirming sentiments in social actions*. Springer Science & Business Media, 2007.

[29] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, pages 57–58, 2011.

[30] Eduard H Hovy. What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis. In *Language Production, Cognition, and the Lexicon*, pages 13–24. Springer, 2015.

[31] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artif. Intell.*, 172(16-17):1897–1916, 2008. doi: 10.1016/j.artint.2008.08.002. URL http://dx.doi.org/10.1016/j.artint.2008.08.002.

[32] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *JASIST*, 60(11):2169–2188, 2009.

[33] Keyuan Jiang and Yujing Zheng. Mining twitter data for potential drug effects. In *Advanced Data Mining and Applications, 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part I*, pages 434–443, 2013.

[34] Jaap Kamps, Maarten Marx, Robert J Mokken, and Marten de Rijke. *Words with attitude*. Citeseer, 2001.

[35] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[36] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[37] Yury Kryvasheyeu, Haohui Chen, Esteban Moro, Pascal Van Hentenryck, and Manuel Cebrián. Performance of social network sensors during hurricane sandy. *PLoS one*, abs/1402.2482, 2014.

[38] Bing Liu. Opinion mining. In *Encyclopedia of Database Systems*, pages 1986–1990. 2009.

[39] James R Martin and Peter R White. *The language of evaluation*. Palgrave Macmillan, 2003.

[40] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*, pages 1155–1158, 2010.

[41] Yelena Mejova, Padmini Srinivasan, and Bob Boynton. GOP primary season on twitter: "popular" political sentiment in social media. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 517–526, 2013.

[42] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[43] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, pages 412–418, 2004.

[44] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *17th International Conference on Intelligent User Interfaces, IUI '12, Lisbon, Portugal, February 14-17, 2012*, pages 189–198, 2012.

[45] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[46] Charles E Osgood. The nature and measurement of meaning. *Psychological bulletin*, 49(3):197, 1952.

[47] Karen OConnor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 924. American Medical Informatics Association, 2014.

[48] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, cs.CL/0205070, 2002.

[49] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of liwc2007, 2007.

[50] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 181–189, 2010.

[51] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

[52] A Platanios, Avrim Blum, and Tom M Mitchell. Estimating accuracy from unlabeled data. In *In Proceedings of UAI*, 2014.

[53] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*, pages 180–185, 2011.

[54] Adam Sadilek, Henry A. Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[55] T. Sakaki, Y. Matsuo, T. Yanagihara, N.P. Chandrasiri, and K. Nawa. Real-time event extraction for driving information from social sensors. In *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012 IEEE International Conference on*, pages 221–226, May 2012. doi: 10.1109/CYBER.2012.6392557.

[56] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):919–931, April 2013. ISSN 1041-4347. doi: 10.1109/TKDE.2012.29.

[57] Suvash Sedhain, Scott Sanner, Lexing Xie, Riley Kidd, Khoi-Nguyen Tran, and Peter Christen. Social affinity filtering: recommendation through fine-grained analysis of user interactions and activities. In *Conference on Online Social Networks, COSN'13, Boston, MA, USA, October 7-8, 2013*, pages 51–62, 2013.

[58] Suvash Sedhain, Scott Sanner, Darius Braziunas, Lexing Xie, and Jordan Christensen. Social collaborative filtering for cold-start recommendations. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, pages 345–348, 2014.

[59] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological online debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics, 2010.

[60] Kate Starbird and Leysia Palen. *Pass it on?: Retweeting in mass emergency.* International Community on Information Systems for Crisis Response and Management, 2010.

[61] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[62] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[63] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, pages 1079–1088, 2010.

[64] Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig H. Martell, and Joseph King. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729, 2012.

[65] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan.

A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

[66] Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, 3(1):6–20, 2003.

[67] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[68] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 625–631, 2005.

[69] Zhiheng Xu and Qing Yang. Analyzing user retweet behavior on twitter. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August 2012*, pages 46–50, 2012.

[70] Yiming Yang, Thomas Pierce, and Jaime G. Carbonell. A study of retrospective and on-line event detection. In *SIGIR '98: Proceedings of the 21st Annual International (ACM) (SIGIR) Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 28–36, 1998.

[71] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter i hope it is not as bad as i fear. *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

[72] Yongzheng Zhang and Marco Pennacchiotti. Predicting purchase behaviors from social media. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 1521–1532, 2013.

[73] Yongzheng Zhang and Marco Pennacchiotti. Recommending branded products from social media. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pages 77–84, 2013.

[74] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *Technical Report TR0620-2011, Rice University and Motorola Mobility*, abs/1106.4300, 2011.