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

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

Abstract

Contents

1 Introduction **2**

2 Use Cases of Social Media Sensor **2**

2.1 Events . . . 2

2.1.1 Trending Topic Detection . . . 3

2.1.2 Physical Event Detection . . . 4

2.2 Sentiments and Opinions . . . 6

2.2.1 Types of Sentiment Analysis . . . 6

2.2.2 Applications of Sentiment Analysis . . . 7

2.3 Preferences and Traits . . . 9

2.3.1 Framework of Preference Prediction . . . 9

2.3.2 Applications of Preference and Trait Prediction . . . 9

3 Supporting Theories **11**

3.1 Sentiment Theories . . . 12

3.1.1 Complex and Simple Sentiment . . . 12

3.2 Social Network Theories . . . 14

3.2.1 Graph Structure . . . 14

3.2.2 Information Diffusion and Cascades . . . 17

4 Discussion of Open Areas **19**

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 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 diverse 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 [79]. 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, and 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 [45]. 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 content.

- **Threads of tweets** Petrovic et al. [57] tried to detect novel events from streams of Twitter posts by forming threads of similar tweets. The minimum similarity distance to an existing tweet represented the novelty score of the tweet. Further, similarity threshold for assigning tweets to threads controlled size of threads. The fastest growing thread in each time interval indicated the news of the event spreading and was outputted as a new event. Becker et al. [8], Ishikawa et al. [35], O'Connor et al. [52], Phuvipadawat and Murata [59] also tried to detect trending topics by clustering and computing similarity degree between words and clusters. Becker et al. [8] additionally considered the classification of tweets as referring to real-world events or not.
- **Wavelet analysis** Weng and Lee [76] applied wavelet analysis to individual words on the frequency based raw signals of 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 [45] detected events by focusing on bursts of keywords whereas Cui et al. [18] used different hashtag properties for this purpose. Zhao et al. [83] and Nichols et al. [50] 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.

Query-based Methods The third category of works focus on the hypothesis that trending topics can be detected by measuring user-defined criteria.

- **Location-dependent** Emphasizing location, Albakour et al. [2] and Sakaki et al. [63] detected local events based on either user query, user location, or both. Albakour et al. employed contents of the tweets and volume of microblogging activity 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 to detect 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.

Network Structure-based Methods The last category of works focus on the hypothesis that trending topics can be detected by studying the network structure of users.

- **Network structure-based** Budak et al. [13] 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 Physical Event Detection

There are many types of physical events that are discussed in social media. This part focuses on research on two important events of this type: natural disasters and health epidemics.

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

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

of works targeting disastrous events on Twitter can be divided into the two following categories:

- **Predictive studies on disaster** Kryvasheyeu et al. [41] 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. [64] used SVM classifier for detecting earthquakes 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 during crisis [17, 68, 71] 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.

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 [16, 22]. 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 [19] and Aramaki et al. [4] 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** García-Herranz et al. [25] use the friendship paradox concept (described in section 3.2.1) for early detection of contagious outbreaks. 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. They claim that this sensor group represents more central individuals and individuals

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

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

at the center of a network are likely to receive a contagion sooner than randomly-chosen members of the population (because central individuals are a smaller number of steps away from the average individual in the network). As a result, García-Herranz et al. [25] argued that this selection process of sensor groups helps in early detection of outbreaks.

- **Hybrid method** Sadilek et al. [62] exploited both content of tweets and structural information of users network. They employed a semi-supervised approach to learn a SVM classifier using n-grams as features in order to detect ill individuals. Then, they estimated physical interaction between healthy and sick people based on co-location and friendship. This enabled them to study the effect of these two factors of social activity (co-location for contact network and friendship for social ties) on public health.

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 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 [43]. 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 [14, 43]. Complex sentiment involves the sentimental reaction of the human to various words across different factors [53, 69, 77]., 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 applications which are discussed in more detail in the following sections.

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

2.2.2 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 [70] which shows the extent and influence social media campaigns hold during political debates and events. However, the question is can social media such as Twitter predict elections?

Researchers have studied social media in order to either investigate and evaluate the relationship of online political sentiment to offline political landscape [6, 51, 70, 73] or to see if online political sentiment can be predictive of actual election results [9, 46]. Methodologies used for these purposes include using textual analysis software (LIWC [56]) [70], classification e.g., Naive Bayes, SVM, Adaboost) [6, 9, 46, 73], or simple statistical methods such as computing sentiment score as the ratio of positive to negative word counts [51]. These methods are based on different sets of features extracted from text such as lexicon-based features [6], the frequency of keywords [51, 70], and with uni-grams being the most commonly used and successful feature [9, 46, 73].

Regarding the predictive power of Twitter, Bermingham and Smeaton [9], Mejova et al. [46] extracted simple sentiment from social media and compared it to actual national polls results. Bermingham and Smeaton [9] claim that social analytics using both volume-based measures and sentiment analysis were predictive of public opinion during the Irish general election. On the other hand, Mejova et al. [46] argue that online sentiment is not predictive of national poll results for US presidential candidates. Tumasjan et al. [70] 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. The analysis of tweets' political sentiment showed close correspondence to the parties' and politicians' political ties claiming that the content of tweets reflect the offline political landscape.

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

Research in this application area targets different products, e.g., movies, laptops, cameras, books, music [20, 55], trends of different brands in social media, and the relationship between the company and customers [26, 36]. 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. [55] used POS-tags and position of words, Dave et al. [20] used other linguistic features e.g., negations and colocation, and Ghiassi et al. [26] extracted emoticons in addition to n-grams.

Moreover, in contrast to [20, 55] who extract simple sentiment, [26, 36] used graded sentiment on a 1 to 5 scale to rank sentiment toward brands. They compared the classification results to scalar rating per product provided in the websites such as Amazon, IMDB, etc. 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.

Stock Market Applications Another application of sentiment analysis can target stock markets with the question of can Twitter predict stock market?

Bollen et al. [12] 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 of future stock market values. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network trained on the basis of past DJIA[5] 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. 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 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. [80] 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 [37, 54] and/or to identify potential drug users [10]. Methodology used in these works is similar to product market research and includes typical classification methods e.g., SVM and Maximum Entropy [10, 37], and manually coded classification with concept extraction and lexicon matching [54] 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 [10, 37, 54], presence and frequency of semantic types of disease or syndrome [37], and textual features e.g., number of hashtags, reply-tags, urls, pronouns [10, 37]. 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 [37].

[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 [5]

2.3 Preferences and Traits

Social media provides sources of data, e.g., people reveal their preferences online, which can be mined. Also, researchers can leverage social networks to improve predictions of preferences and traits. There are two types of preference learning problems on social media: personalized, and collaborative. The first is where there is only a single user and many items. Usually, researchers use product description as features of the item in order to predict preferences and the predictions are shown as ranking of items [29]. The second case is when there are multiple users and multiple items. This scenario is often called collaborative filtering [34]. Learning the user's preferences can help in understanding what users prefer to buy, who 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. The most important target users of this procedure are marketing companies and political parties.

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 $\mathit{Rating}(U_1, X_2) = 3$. 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 for preference prediction:

- Content-based: methods based on features extracted from the content of posts by employing simple linear regression, classification, or data mining approaches
- Social-based methods: methods dependent on the links and interaction between users (share, comment, tag, mention, like, retweet). These methods are based on homophily, the theory that individuals with similar characteristics or interests are more likely to form social ties [1]
- 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

specific instances of these methods are outlined in next section.

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[40] 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 Pearson's correlation[60]. Methods used by [40] and [60] are both interaction-based. Numeric variables such as age or intelligence were predicted using a linear regression model, whereas dichotomous variables such as gender or sexual orientation were predicted using logistic regression. Kosinski et al. [40] used Facebook likes as the only feature, while Quercia et al. [60] 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 [81, 82], numeric real-values showing the preferences for each item [66], or binary values on whether the user would like an item or not [65]. They used different methodologies such as simple popularity methods [81], linear regression [65, 81, 82], simple classifiers (Naive Bayes, SVM, logistic regression) [65, 81], or collaborative filtering methods based on matrix factorization [66]. Zhang and Pennacchiotti [81, 82] 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. [65] 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 [27, 28], classifying stances on political debates concerning topics of health care, gay rights, gun rights, ... [67, 72], 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. [28] applies collaborative filtering based on probabilistic matrix factorization.
- The non-collaborative works either use simple data mining and statistical approaches [27], homophily measure between users and their followers/followees using similarity metrics [1], or classification methods [67, 72]. To apply these methods, researchers extracted features including sentiment features [67, 72], and structure-based features (network of users on following each other) [1, 27].

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.

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

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

important for understanding and predicting information diffusion in Twitter. To this end, various works have been proposed. In the following, these works are categorized based on two different main goals:

1. **Predict if a tweet will be retweeted in future and provide retweet count** [15, 58, 78]: All of these works use classification-based approaches using tweet-based and author-based features. However, Can et al. [15] took advantage of visual cues from images linked in the tweets, and Xu and Yang [78] employed social-based features in addition to tweet author-based features. Different from the other two works, Xu and Yang [78] performed the analysis from the perspective of individual users. Petrovic et al. [58] worked on retweet prediction of real-time tweeting with online learning algorithms and claimed that performance is dominated by social features, but that tweet features add a substantial boost.

2. **Rank tweets based on retweeting probability or category** [24, 31]: Works in this category focus on finding important tweets by analyzing propagation of tweets through retweeting. Feng and Wang [24] used author-based and interaction-based features in addition to tweet-based features to build a graph in order to model retweet behavior. They designed a model that learns latent biases for each node based on the underlying graph. Their model is based on the notion that tweet history reveals user's personal preference. Hong et al. [31] formulated ranking tweets into a two-step classification problem by investigating features based on content, temporal information, users, and topological features of user's social graph. The first classifier predicts whether a tweet will be retweeted, while the second classifier predicts volume range of future retweets for a new messge.

These studies showed that temporal features have a stronger effect on messages with low and medium volume of retweets compared to highly popular messages, and user activity features can further improve the performance marginally. Also, Hong et al. claimed that *degree distribution* and *retweet before* contribute greatly to retweet behavior. Feng and Wang [24] mentioned that importance of a tweet varies from user to user, and considering publisher's authority and tweet's quality alone is not enough, personalization plays an important role in the retweet behavior.

Feature types used in the above methods are shown in more detail in table 1.

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.

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	color histograms
Topological	Page-rank, degree distribution, local clustering coefficient, reciprocal links

Table 1: List of features used in retweet prediction

3.1 Sentiment Theories

Sentiment theories cover the characteristics of text, necessary for determining the attitude of the author. Attitude can be based on (1) author's judgment [77], (2) affective or emotional state [53], or (3) the intended emotion the author wanted to convey [32]. (1) gets the emotion from author's point of view on a subject, (2) conveys the state of the author at the time of writing, and (3) is the emotional effect that the author was trying to convey to the reader. Considering the example of using humor in regards to product review e.g., "It could not be any better, it broke in two days.", it becomes clear that research in sentiment analysis should investigate multiple dimensions.

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 for 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 [14], outlined as follows:

Sentimental Reaction to Various Words Osgood [53], in a study of text polarity showed human's sentimental reaction to various words across eight dimensions, for example, three dimensions are the following:

- Evaluation (positive or negative)
- Potency (strong or weak)

Dimensions	**Positive side**	**Negative side**
Evaluation	nice, sweet, heavenly, good, mild, happy, fine, clean	awful, sour, hellish, bad, harsh, sad, course, dirty
Potency	big, powerful, deep, strong, high, long, full, many	little, powerless, shallow, weak, low, short, empty, few
Activity	fast, noisy, young, alive, known, burning, active, light	slow, quiet, old, dead, unknown, freezing, inactive, dark

Table 2: Positive and negative side of dimensions

- Activity (active or passive)

Each aspect is characterized by a variety of contrasts. Characterizations of the positive and negative side of each dimension are shown in table 2 [30]:

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 [77]. In general, appraisal theory is an analysis of how a writer values people and things within the text that he/she produces [44]. It studies different types of evaluative language that can occur and represents three grammatical systems comprising appraisal [11]:

- Attitude: tools that an author uses to directly express his approval or disapproval of something, further divided 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, etc.
- Graduation: resources which an author uses to convey the strength of that approval or disapproval such as very, reasonably, ...

Hence, this theory and Osgood [53]'s theory (with three dimensions of evaluation, potency, and activity) 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 [38, 48] for classifying words.

Psycho-Linguistic Theories The third theory focuses more on the psychological aspect of language and how people with different psychological backgrounds use words, also known as LIWC dictionary [69]. 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 [69] provided 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. They argued that it is possible to relate daily word use e.g., nouns, adjectives, verb tenses, etc. to a broad array of real-world behaviors and different emotional indicators e.g., emotional state, social relationships, thinking style, etc. For example:

- positive political ads used more present and future tense verbs, or people used past tense more frequently in discussing a disclosed event
- higher-status individuals have greater use of first-person plural and ask fewer questions compared with lower-ranked ones
- deceptive statements use more negative emotions, more motion words (e.g., arrive, car, go), fewer exclusion words, less first-person singular, higher total word count, and more sense words

3.2 Social Network Theories

Every Social Media has an underlying social network structure. Studying the structure of this network and the elements of information diffusion as underlying parts of many applications, is important. This section is devoted to social network related theories that correlate with discussed applications in the first part of the survey.

3.2.1 Graph Structure

Social Networks are comprised from graphs with special properties owing to their sociological origins. Different types of graph structures have been introduced through history. Here, we provide studies on how social network graphs are generated, how information flows through social networks, and how different users play structurally distinct roles. Moreover, we discuss 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 [21, 74]. 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, thus bridging 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 represent 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 C, D, and E, such that if a strong tie exists between $C - D$ and $C - E$, there is a weak or strong tie between $D - E$.
- **Centrality** Characterizes the importance of nodes (individuals) by measuring information brokerage in social networks. The degree centrality (in undirected networks, individuals with higher connections have more risk of catching whatever is flowing through the network such as information), closeness centrality (in undirected networks, measures the total distance from all other individuals), betweenness centrality (in undirected networks, a measure for quantifying the control of an individual on the communication between other individuals in a social network), Katz centrality or PageRank(in directed networks, a measure for estimating importance of an individual by counting the number and quality of links to her/him). Some of these measures are shown in Figure 1.

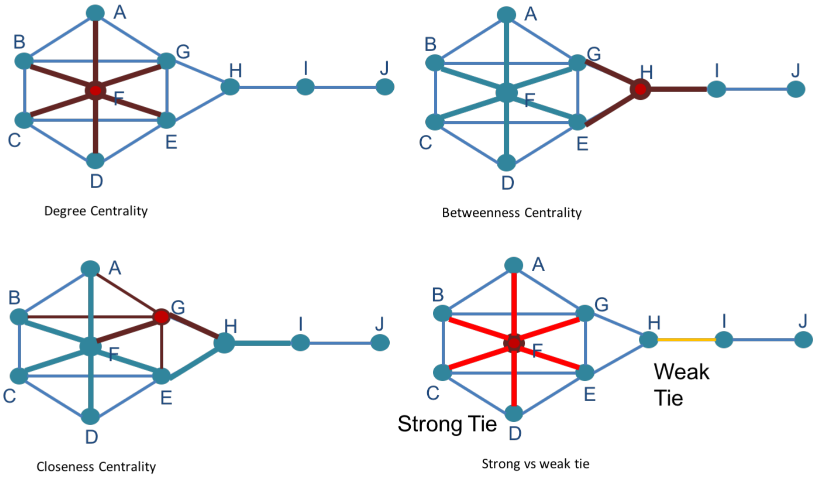

Figure 1: Different centrality criteria and Weak vs. Strong ties

Graph Generation Models Historically, four different graph models have been studied which are discussed here. It is important to become familiar with these models in order to better understand the differences and the unique attributes that only Social Networks represent, such as preferential attachment which refers to the observation that in growing networks (over time), the probability that an edge is added to a node with d neighbors is proportional to d. However, what are the proposed models and what salient properties of social networks do they represent?

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

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

3. Scale-free networks: graphs characterized by a highly heterogeneous degree distribution, which follows a ”power-law” distribution
4. The Barabasi-Albert model: the first network with a power-law distribution which are random scale-free networks generated using preferential attachment mechanism.

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 (described above)

Random networks, also known as Erdos-Renyi networks, are an entirely random network based on a probability p of connecting nodes. These networks have short path length and independent edges. The concept of small-world networks was introduced by Watts-Strogatz in which most nodes can be reached from every other in a small number of steps (following the six degrees of separation theory). Social networks are not purely random graphs or Watts-Strogatz graphs since they represent both preferential attachment and small world behavior.

Unlike the last two static structures, scale-free networks are dynamically formed by continuous addition of new nodes to the network. The two main ingredients for self-organization of a network in a scale-free structure are 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. These networks have smaller average path length compared to random graphs and small-world networks [74]. Albert and Barabási [3] introduced an algorithm for generating a scale-free network with power-law distribution and having two ingredients of growth and preferential attachment.

Over the years, researchers have uncovered scale-free structures in some social networks such as sexual relationships among people in Sweden[9], network of people connected by email, network of scientific papers connected by citations, ...

An important corollary of graph structures is discussed next.

Friendship Paradox

The concept of friendship paradox is derived from graph generation models and their properties. Feld [23] introduced the concept of *friendship paradox*. Using general mathematical properties of social networks, he showed that on average most people have

[9] Albert-Laszlo Barabasi and Eric Bonabeau

fewer friends than their friends have and he called this the ”Friendship Paradox”. 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 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 number of friendship relationships (degree) of people in the graph. And the average number of friends that a typical friend of a person has, was modeled by choosing uniformly at random an edge (a pair of friends) and an endpoint of that edge (one of the friends), followed by calculating the degree of the selected endpoint again. By considering properties of variance and mean of degrees (friendship relationships) and this modeling, Feld formally proved that, on average, your friends have more friends than you do [23].

This implies that friends of a randomly selected person are likely to have higher than average centrality which is an important concept in various applications such as in case of epidemics and outbreak detection as discussed in 2.1.2.

3.2.2 Information Diffusion and Cascades

Social networks have emerged as a critical factor in information dissemination, search, marketing, expertise and influence discovery. 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 pattern of epidemic spreading. Information cascade occurs when a person observes the actions of others and then decides to engage in the same act based on pay-off benefits of one strategy or the other. Epidemic spreading though is when the process of contagion is complex and unobservable and doesn’t involve decision making by users.

Epidemic spread In classic epidemiology individuals have an equal chance of contact i.e. homogeneous contact network. However, this was determined to be unrealistic. In response, Newman [49] introduced an underlying contact networks model [49]; Contact networks represent individuals as a nodes and contacts 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 [47]. Epidemic spread models have been used in social media for studying the effects of information going viral, in different applications such as internet memes, news, etc.

The terminology for epidemiological models include the following variables:

- S: Susceptibles
- E: Exposed individuals in the latent period
- I: Infectives
- R: Recovered with immunity

- β: Contact rate

The three variables $S(t)$, $E(t)$, $R(t)$, $I(t)$ represent the number of individuals with the specified state at the time t and β is a real valued variable showing the contact rate or the probability of contagion after contact per unit of time.

Similar to this, social contagion phenomena refers to various processes that depend on the individual propensity to adopt and diffuse knowledge, ideas, and information. In social contagion we have similar concepts:

- S: an individual who has not learned new information
- I: the spreader of the information
- R: aware of information, but no longer spreading it

Famous epidemiological models include:

1. SI: Susceptible-Infected [49]
2. SIR: Susceptible-Infected-Removed [39]
3. SIS: Susceptible-Infected-Susceptible
4. SEIR: Susceptible-Exposed-Infected-Removed

These models show potential stages individuals would go through and they model number of individuals in each stage as random variables. In general, patterns of epidemic spread depend on a disease's contagiousness β.

Studying the dynamics of epidemics on graphs, suggested the existence of an epidemic threshold above which epidemics spread to a significant fraction of the graph [75]. This is of high importance in studying how news, video, and opinion become viral [7, 42].

Diffusion Models Building on epidemic models, researchers could define information diffusion properties. Given a social network and estimates of reciprocal influence, viral marketing, also known as the influence maximization problem, is defined to target the most influential users in the network in order to activate a chain-reaction of influence and eventually influence largest potential number of users in the network [61]. There are studies demonstrating how a model of the diffusion of information can be used to study information cascades on social media such as Twitter that are in response to an actual crisis event [17, 33].

Two basic classes of diffusion models exist: Linear Threshold Model and Independent Cascade Model. These models represent a social network as a directed graph with a binary variable for infection associated to each node (person). 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 for all time. The activation attempt succeeds with a certain probability related to those two nodes.

4 Discussion of Open Areas

While there is rich work covering use of social media sensors, we now identify a few critical gaps in the social media literature:

- Learning social media sensors by refining current learning paradigms to improve their usability for detection of general topical tweets. A large body of work in detection of these tasks define use of sensors, but rely on ad-hoc methods based on data mining, simple statistics, or off-the-shelf classifiers to select sensors. In contrast, we intend to learn tunable sensors specific for each topic with the ability to learn on a smaller subset of data and generalize to large amount of data.
- Semi-supervised learning methods: one of the key challenges of learning social media sensors is providing the appropriate training data for the model to learn. Considering the scale of the dataset and the diversity of information needs on social media, it is simply not possible (even with crowdsourcing) to manually annotate millions of tweets/posts and decide whether they are related to a specific topic as an example. Most of the current works either focus on smaller datasets, which defeats the purpose of using social media as a huge source of information, or use very specific targets in order to automatically label all the positive cases by use of very specific hashtags/terms. This shows a gap for moving to semi-supervised learning methods to leverage the outputs of multiple independent classifier for final decisions.

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, 2013.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *CoRR*, cond-mat/0106096, 2001.

[4] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: De-

tecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 2011.

[5] 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.

[6] 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.

[7] Christian Bauckhage, Fabian Hadiji, and Kristian Kersting. How viral are viral videos? In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[8] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[9] 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.

[10] 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.

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

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

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

[14] Sarah Bull. Thesis: Automatic parody detection in sentiment analysis. 2010.

[15] 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.

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

[17] 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.

[18] 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.

[19] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, 2010.

[20] 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, 2003.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] Shengbo Guo and Scott Sanner. Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 289–296, 2010.

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

[31] 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.

[32] 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.

[33] Cindy Hui, Yulia Tyshchuk, William A. Wallace, Malik Magdon-Ismail, and Mark K. Goldberg. Information cascades in social media in response to a crisis: a preliminary model and a case study. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, pages 653–656, 2012.

[34] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artif. Intell.*, 172(16-17):1897–1916, 2008.

[35] S. Ishikawa, Y. Arakawa, S. Tagashira, and A. Fukuda. Hot topic detection in local areas using Twitter and wikipedia. In *ARCS Workshops (ARCS), 2012*, pages 1–5, Feb 2012.

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

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] Kristina Lerman, Rumi Ghosh, and Tawan Surachawala. Social contagion: An empirical study of information spread on digg and Twitter follower graphs. *arXiv preprint arXiv:1202.3162*, 2012.

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

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

[45] 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*, pages 1155–1158, 2010.

[46] 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.

[47] 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.

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

[49] Mark E. J. Newman. Networks - an introduction (2010, oxford university press.). *Artificial Life*, 18:241–242, 2012.

[50] 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.

[51] 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.

[52] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

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

[54] 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.

[55] 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.

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

[57] 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.

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

[59] Swit Phuvipadawat and Tsuyoshi Murata. Breaking news detection and tracking in Twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops, Toronto, Canada, August 31 - September 3, 2010*, pages 120–123, 2010.

[60] 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*, pages 180–185, 2011.

[61] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM, 2002.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

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

[75] Yang Wang, D. Chakrabarti, Chenxi Wang, and C. Faloutsos. Epidemic spreading in real networks: an eigenvalue viewpoint. In *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*, pages 25–34, 2003.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through Twitter. *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

[81] 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.

[82] 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.

[83] 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.