AN ABSTRACT OF THE THESIS OF

Zahra Iman for the degree of Master of Science in Computer Science presented on May 25, 2016.

Title: Learning Topical Social Media Sensors for Twitter

Abstract approved: _____

Scott P. Sanner

Social media sources such as Twitter represent a massively distributed social sensor over diverse topics ranging from social and political events to entertainment and sports news. However, due to the overwhelming volume of content, it can be difficult to identify novel and significant content within a broad topic in a timely fashion. To this end, this thesis proposes a scalable and practical method to automatically construct social sensors for generic topics. The concept of using social media as a sensor for detection of events and news has been proposed in the literature. However, we argue that most of these works do not focus on targeted content detection or they use very basic methods for collecting the topical data for further analysis. This demonstrates a gap in the use of social media as a sensor for high-quality topical content detection that we aim to address via machine learning. In this thesis, given minimal supervised training content from a user, we learn to identify topical tweets from millions of features capturing content, user and social interactions on Twitter. On a corpus of over 800 million English Tweets collected from the Twitter streaming API during 2013 and 2014 and learning for 10 diverse topics, we empirically show that our learned social sensor automatically generalizes to unseen future content with high ranking and precision scores. Furthermore, we provide an extensive analysis of features and feature types across different topics that reveals, for example, that (1) largely independent of topic, simple terms are the most informative feature followed by location features and that (2) the number of unique hashtags and tweets by a user correlates more with their informativeness than their follower or friend count. In summary, this work provides a novel, effective, and efficient way to learn topical social sensors requiring minimal user curation effort and offering strong generalization performance for identifying future topical content.

Learning Topical Social Media Sensors for Twitter

by

Zahra Iman

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented May 25, 2016
Commencement June 2016

Master of Science thesis of Zahra Iman presented on May 25, 2016.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Zahra Iman, Author

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Dr. Scott Sanner for his support and guidance throughout this thesis, I have learned everything I know about research from him. Thanks go out to my fellow researchers, Reda Bouadjenek and Dan Nguyen for their contributions to this thesis. I would also like to thank Matteo Smullin for supporting me every step of the way. Also, I would like to thank my family who have supported me through this entire process by keeping me harmonious and helping me put the pieces together. I will be forever grateful for your love.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure | Page

LIST OF TABLES

LIST OF APPENDIX FIGURES

Figure Page

LIST OF APPENDIX TABLES

Table Page

Chapter 1: Introduction

1.1 Motivation

Social networks such as Twitter and Facebook have millions of users from diverse backgrounds, who tweet and post about localized urban issues such as potholes, car accidents, traffic jams, and public transport as well as family, education, health and sports events. They also tweet and post about wider-ranging national and global issues such as natural disasters, epidemics, and politics. Viewed as a whole, social media provides a rich perspective on humanity ranging from minor, localized personal observations to topics of global concern.

The content of these posts could be of importance to government agencies such as U.S. FEMA (Federal Emergency Management Agency) and the U.S. CDC (Centers for Disease Control and Prevention) to prevent casualties in case of disasters. It could help news websites with important content on sport events, celebrities, social issues, etc. It could be useful to political parties to know the users ideas on certain issues. It could also be of great help to companies to know what are the preferences of their users, what do customers complain about their product, etc. In order to deal with global challenges, improve emergency management, and achieve a higher quality of life, there is a need to capture and make use of this massive amount of information.

However, social media sites present a conundrum for users. On one hand these sources contain a vast amount of novel and topical content that challenge traditional news media sources in terms of their timeliness and diversity. Yet on the other hand they also contain a vast amount of spam and otherwise low-value content for most users' information needs where filtering out irrelevant content is extremely time-consuming. As an example, Twitter has 302 million active users and they send 500 million tweets per day[1]. However, Twitter's search does not facilitate targeted information extraction covering individualized information needs. Currently, Twitter only provides search through only Boolean retrieval with temporal ranking [2]. This search method

[1] http://www.marketingprofs.com/chirp/2015/28385/how-the-world-uses-twitter-infographic
[2] https://blog.twitter.com/2011/the-engineering-behind-twitter-s-new-search-experience

critically fails to capture the underlying query intent when non-exact matches or more informative historical content may be more useful to the user than the most recent, exact matches. The lack of smart search methods represents a clear and present need for a more intelligent method to search for relevant topical content from massive numbers of posts.

1.2 Social Media Sensors

This thesis explores the use of social media as a sensor. A Social Media sensor as defined in the literature "collects, processes, and aggregates big streams of social media data and multimedia to discover trends, events, influencers, and interesting media content"[3]. For example we could learn a classifier to predict whether a tweet is related to *Natural Disaster* and thus building a "social sensor" for *Natural Disaster*. Existing literature on use of social media as a sensor covers:

- Designing/developing social media sensors for detection of a specific topic such as earthquake or flu [73, 19, 71]
- Designing/developing social media sensors for trending topic detection [66, 53, 61]
- Designing/developing social media sensors for sentiment detection [79, 9, 12, 10]
- Designing/developing social media sensors for preference detection and tweet recommendations [43, 91, 15, 25]
- Designing/developing social media sensors for tracking general topics [47, 88, 49]

While all of these works try to take advantage of concept of social media sensors, majority of the literature works are either too narrow and highly specific to a certain topic, focus only on ad-hoc methods and not taking advantage of learning, not targeted on any topic, they use very basic methods for collection of their data for further analysis, or like [47, 88, 49] cover part of the overall framework that we propose and omit some potentially important issues such as ranking and validation; we present a comprehensive literature review of social sensors and contrast this with our proposed work in Chapter 2. Hence, the goal of this thesis is to provide users with a novel method to build a more flexible search tool for Twitter. Our work combines, extends, and provides a longitudinal study of learning general topical social sensors.

[3] http://www.socialsensor.eu

1.3 Contributions

The contribution in this work falls into two main categories.

1. **Supervised learning of topical social sensors**

   This contribution answers the following questions:

   (a) How do we learn a method for extracting informative content on generic topics from millions of features and small sets of examples while minimizing required user input?

   (b) Which classification method(s) would have higher precision and would generalize to broader related content?

   To this end, we provide a novel supervised method for learning high precision topical social sensors. On the corpus of over 800 million Tweets and covering 10 diverse topics ranging from ”social issues” to ”celebrity deaths” to the ”Iran nuclear deal”, we empirically show that two simple and efficiently trainable methods — logistic regression and Naïve Bayes — are capable of learning users information needs and generalize well to unseen future topical content (including content with no hashtags) in terms of their mean average precision (MAP) and Precision@n for a range of n.

2. **Longitudinal study on the features and their attribution**

   This contribution uncovers insights from the dataset and answers the following questions:

   (a) What are the best features for learning social sensors and do they differ by topic?

   (b) For each feature type, do any attributes correlate with importance? For example, are users with higher number of followers more important? What is the relationship between number of tweets a hashtag has been used in and the importance of the hashtag?

   To this end, we deliver a longitudinal and comprehensive study on the features extracted from Twitter and their attributes as they relate to retrieving relevant social content. Overall, feature analysis in this work shows that in general learning methods may be more effective than manual engineering for building topical social sensors. We draw a number

of important insights from our analysis, with two of the more surprising insights being the following:

(a) Despite the fact that training labels were derived from hashtags, we found out that *terms* and *locations* are among the most useful features.

(b) While one might hypothesize that users with higher follower and friends counts are more informative, we found out that the number of unique hashtags and tweets by a user correlates more with their informativeness than their follower or friend count.

In summary, this work fills a major gap in event detection and tracking from social media on identifying emerging topics from long-running themes with minimal user supervision. Our results suggest that these sensors generalize well to unseen future topical content and provide a novel paradigm for the extraction of high-precision content from social media.

1.4 Outline of Thesis

The thesis is organized as follows: In Chapter 2 the related work on using social media as a sensor is reviewed. Chapter 3.2 describes the dataset used in this work. Chapter 3 gives a detailed description of the proposed framework for learning supervised topical social media sensors and the results. Chapter 4 provides the feature analysis on Twitter. In Chapter 5 we conclude by highlighting the challenges and providing suggestions for future research in using social media as a sensor for latent topic detection.

Chapter 2: Literature Review

2.1 Introduction

Prior to this work, a vast amount of literature has examined the use of social media as a sensor where the majority of this work can be organized along four major uses cases: event detection, tracking broad topics, sentiment analysis, and preference learning. We cover each of these use cases in detail in this section and identify shortcomings and critical gaps that prevent them from learning topical social sensors as proposed in this thesis.

2.2 Events

One of the major use cases of social media sensors are detecting i.e. sensing 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 [89]. 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.

2.2.1 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 [53]. 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. [66] 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. Ishikawa et al. [38], Becker et al. [8], Phuvipadawat and Murata [68], O'Connor et al. [61] 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 [85] 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.

Considering that this thesis focuses on generic and broad topic detection, the main issue with clustering based methods is that they are unsupervised and we want to use supervised learning of topical information detection for generic topics.

Burst-Detection 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.

Mathioudakis and Koudas [53] detected events by focusing on bursts of keywords whereas Cui et al. [18] used different hashtag properties for this purpose. Zhao et al. [93] and Nichols et al. [59] 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. Emphasizing location, Albakour et al. [2] and Sakaki et al. [72] detected events targeted on user's query using burst patterns. 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.

Burst-detection methods focus on temporal patterns of specific sets of keywords and therefore suffer from (1) not being targeted, (2) overlooking topical content that is not trendy i.e. topical content that does not follow the same specific properties of terms such as bursts, and (3) lack of learning methods makes it almost impossible to generalize to future unseen topical keywords.

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.

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.

The issue with network structure-based methods, similar to other methods for finding trending

topics, is that they are also not targeted to a specific topic. Also, considering that our dataset contains over 95 millions of users it would be almost impossible to access the social network of all the users – especially since Twitter streaming API gives random access to only 1% of the tweets.

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

- **Predictive studies on disaster** Focusing on hurricane Sandy, Kryvasheyeu et al. [44] 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] explained in more details in Appendix A). Our topical social sensors represent a superset of user-based sensors discussed in this work since our work includes user-based features when the predictor learns to use them. However, as shown in our feature analysis in Chapter 4, user-based features are among the least informative feature types for our topical social sensors suggesting that general social sensors benefit from a wide variety of features well beyond those of author features alone.

  Focusing on earthquake, Sakaki et al. [73] 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. While this work is of high value for detection

[1] http://mashable.com/2009/08/12/japan-earthquake/
[2] On average, most people have fewer friends than their friends have

of earthquake events, it is very narrow and highly specific to detection of *earthquake* only and thus the method of this work could not be adopted to unknown or general topics.

- **Descriptive studies on disaster** Related works discuss the behavior of Twitter users during crisis [80, 17, 77] 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 Building a social sensor to detect health epidemic outbreaks is one of the most important use cases of social media sensors. 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 [22, 16]. Target users of this event detection include government agencies such as the U.S. 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. While these works provide important methodology for using social media to sense a topic that could have been used for detection of general topics, they focus on a set of selected keywords/-textual feature related to *Flu* as the features for their classifier. This makes their classifiers highly specific to the sole topic of *Flu detection*.

- **Structure-based method** García-Herranz et al. [27] use the friendship paradox concept (described in section A.0.2.1) for early detection of contagious outbreaks. They provided

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

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 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. [27] argued that this selection process of sensor groups helps in early detection of outbreaks. While the methodology for choosing sensors provided early detection of outbreaks in this case, it is a heuristic method and lacks learning. In addition, our feature analysis in Chapter 4 shows that user-based features are among the least informative features for general topical social sensors. This suggesting that García-Herranz et al. [27]'s method may not readily extend to learning high-fidelity social sensors for general topics.

- **Hybrid method** Sadilek et al. [71] 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. Using learning to detect ill individuals from social media posts is very similar to the topic of this thesis, however [71] does not extend this methodology to the detection topics from the perspective of general topical content detection.

2.2.3 Summary

In this section, we presented existing works on the use of social media for sensing trending topics and physical events. We argued that the main limitation to this important and very active area of research is that (1) trending topic detection is intrinsically unsupervised and not intended to detect targeted topics, and (2) while the Physical event detection methods have the potential of providing high precision event detection, they are highly specific to the target event and do not easily generalize to learn arbitrary event-based or topic-based social sensors as provided in this work. In addition, the methods discussed for Physical Event Detection use very primitive

methods for curating their data. For example, Culotta [19] and Sakaki et al. [73] both use keywords like ”flu”, ”quake”, ”earthquake”, and ”shake” to get their training and testing data. A more robust and complete method for detection of topical content could help the performance and accuracy of these methods as well. In contrast, the work in this thesis is based on supervised learning of a specific topical social sensor derived from the topical set of hashtags provided by the user.

2.3 Tracking General Topics

This section of existing works discusses use of social media sensors for detecting and tracking general topics such as ”Baseball”, ”Fashion”, etc. There are three works fitting this category. Here, we cover these three works by breaking down the general topic learning problem into its components:

- **Collection of labeled data**: Lin et al. [47] collected seven days of unfiltered and unsampled data from Twitter and labeled them based on one hashtag selected for each of the selected 10 topic. Magdy and Elsayed [49] collected four days of data and labeled the tweets based on a user-defined query for each of the three selected topics. Yang et al. [88] took advantage of a multi-step process for collecting labeled data by streaming tweets through a set of topic priors including obvious Twitter accounts of the topic, named entities, and URLs followed by applying co-training based data cleaning algorithm. We build our work on Lin et al. [47]’s work, however we choose a set of hashtags for each topic instead of a single hashtag to (1) cover as much topical content as possible, and (2) to evaluate the generalization of our work on picking future unseen topical hashtags (and thus future unseen topical content).

- **Design of classifier**: Lin et al. [47] leveraged language models (LM) to train models using unigrams and bigrams while Magdy and Elsayed [49] extracted hashtags, unigrams, users and mentions separately as features and applies SVM classifier for binary classification of tweets. Different from the first two, Yang et al. [88] extract hashed unigram frequency and hashed byte four-gram as features and defined the problem as topic modeling of tweets. IN line with these works, we extract hashtags, mentions, unigrams, users as features. We add locations as another feature which we show in Chapter 4 that is location is the second

most important feature for detection of topical content.

- **Training of classifier**: Lin et al. [47] applied LMs by computing the probability of each unigram/bigram based on its usage over a long period of time which is later smoothed based on the recent usage provided by the wordcount within a specific history window. Magdy and Elsayed [49]'s trained off-the-shelf Binary SVM classifer to detect topical content from Twitter. Yang et al. [88] decided on the final topic of the tweet through weighted majority voting of various predictors including (1) a web-page classifier that classifies the tweet based on the embedded URL in the tweet, (2) a tweet classifier that classifies the tweet based on it's content. We apply supervised learning for our research as well, however we provide a novel framework for learning in terms of splitting the data and hashtags as topical proxies, that would ensure matching generalization to future unseen content.

- **Evaluation of results**: Lin et al. [47] reported the results of their LMs for detection of topical content by presenting precision-recall curve of each topic stating that unigram LMs are much more effective than bigram LMs. Magdy and Elsayed [49] reported their results in the form of mean precision and recall for each day and topic and [47]'s results show that when considering two applied classifiers, multinomial logistic regression and one-vs-all logistic regression, the latter works better. Unfortunately, none of them compared the results to other works which makes it hard for us to discuss which one provided better results. We, on the other hand, provide ranking in addition to correct prediction and thus report average precision and precision@n (for a range of n) for each topic as well as mean average precision over all topics. We remark that our train, validation, test framework supports evaluation of the methodology for generalization to future unseen content (More details are provided on Chapter 3).

While these works provide a good basis for this thesis, there are many fine-grain but important differences between previous works and this thesis with the most important ones being:

- We analyzed long-term sensor performance on detecting topical content over two years of Twitter data and across a variety of topics.

- We provide a novel and clear framework for splitting hashtags to train, validation and test in a way ensuring generalization to future unseen content.

- We present ranking in addition to correct classification while none of the other works provide ranking.
- We deliver a comprehensive longitudinal study on features and their attributes over two years of tweets that supports our insights for learning and relevance of features to topicality while these works had little or none analysis over their features.
- We extract *Location* as one of the features which none of these works do and as we show in our feature analysis in Chapter 4, *Location* is the second most important feature beating even hashtags in terms of correlation with topicality.

2.4 Sentiments and Opinions

Sentiment and opinion mining from social media constitutes an important class of social sensors with many important use cases such as marketing companies, government agencies, and individuals are concerned with what users think about them/their products. Considering that the literature on social sentiment and opinion mining can inform the general design of social sensors, we cover related existing works on social sentiment and opinion mining in this section.

2.4.1 Types of Sentiment Analysis

Sentiment analysis, also known as opinion mining, is defined as analysis of text based on expressed sentiments by users. Before we proceed to discuss specific social sentiment sensor methods, we first pause to discuss the diverse output space possible with sentiment-based social sensors. Two major output spaces of sentiment analysis are the following:

Subjective vs. Objective sentiment At the top level of analysis, sentiment can be classified as subjective or objective [48]. 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 [48, 14]. Complex sentiment involves the sentimental reaction of the human to various

words across different factors [86, 62, 78]., 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 with different detection methodologies which are discussed in more detail in the following sections.

2.4.2 Applications of Sentiment Analysis

With the various definitions of sentiment in hand from the previous section, we now seek to use these in the application of social sentiment detection tools to a wide range of applications including political, product market, stock market, and pharmacovigilance applications that we discuss next.

Political Applications Here, we explore works on using social sensors for detecting, i.e. sensing, people's opinions on political issues or political parties. 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 [79] 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 [79, 6, 60, 82] or to see if online political sentiment can be predictive of actual election results [55, 9]. Methodologies used for these purposes include using textual analysis software (LIWC [65]) [79], classification e.g., Naive Bayes, SVM, Adaboost) [55, 82, 9, 6], or simple statistical methods such as computing sentiment score as the ratio of positive to negative word counts [60]. These methods are based on different sets of features extracted from text such as lexicon-based features [6], the frequency of keywords [79, 60], and with uni-grams being the most commonly used and successful feature [55, 82, 9]. The main issue with these set of works is that they are narrow and very focused on the specific political applications. Also, they use very primitive method for gathering their data e.g., collecting data by searching Twitter using keywords e.g., politicians names. Hence,

our methodology for high-precision topical content detection for generic topics could be quite helpful for gathering most informative data for further analysis in this application.

Regarding the predictive power of Twitter, Bermingham and Smeaton [9], Mejova et al. [55] 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. [55] argue that online sentiment is not predictive of national poll results for US presidential candidates. Tumasjan et al. [79] 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. All of these methods mainly focus on keywords and apply ad-hoc methods for sentiment detection which limits their future extension. Rather, the focus of learning topical social sensors is to learn to predict for general topics in a way that generalizes beyond existing labeled topical content to novel future topical content.

Product Market Applications Just as with political applications, social sensors have seen intensive use in detecting opinions on different products. 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, 64], trends of different brands in social media, and the relationship between the company and customers [39, 28]. 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. [64] used POS-tags and position of words, Dave et al. [20] used other linguistic features e.g., negations and colocation, and Ghiassi et al. [28] extracted emoticons in addition to n-grams.

Moreover, in contrast to [64, 20] who extract simple sentiment, [39, 28] 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.

While these methods take advantage of supervised learning to detect sentiment for social media, some of their feature selection such as POS-tags, colocations and their classifiers are very context specific to the topic of product market and not generalizable to other topics.

Stock Market Applications Like political and product market applications, social sensors can also be applied to predict stock market movements. The following works want to answer whether Twitter can 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[4] 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. [90] also showed that Twitter posts can be used to predict market indices. These works present an interesting, complex view of sentiment, however no one has analyzed complex sentiment features in topical social sensors and it remains an unexplored area of research.

Pharmacovigilance Applications Instead of using social sensor for monetary gain, they also have the ability to proactively detect i.e. sense potential Adverse Drug Reactions (ADR) in a population.

Researchers have investigated Twitter posts looking for potential signs of ADR [40, 63] 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 [40, 10], and manually coded classification with concept extraction and lexicon matching [63] in

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

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[5] concerning mention of ADRs [63, 40, 10], presence and frequency of semantic types of disease or syndrome [40], and textual features e.g., number of hashtags, reply-tags, urls, pronouns [40, 10]. 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 [40].

Both works by Jiang and Zheng [40] and Bian et al. [10] provided important work for pharmacovigilance application using supervised learning for detection of ADR. However their use of semantic features from MetaMap, which is focused on medical keywords, makes the learning focused on the specific application and not applicable to generic topics.

2.4.3 Summary

This section explored use of social media sensors for detecting sentiment. We presented the existing works on four various applications: political, product market, stock market, and pharmacovigilance applications. The majority of these works used supervised learning methods for social sentiment detection, which is in line with our thesis's focus on employing supervised learning methods for learning general topical social sensors. We note that while some of the features extracted were targeted on a specific application, such as features extracted from MetaMap, other set of features such as analyzing complex sentiment features could provide a useful hypothesis space of features for future research. Another set of works [63, 6] used ad-hoc methods such as lexion-based approaches, which are too specific to the task of sentiment analysis. In the end, it is important to note that typically there is a lot of labeled sentiment data e.g., online reviews, whereas for generic topic sensors this is not the case. Furthermore sentiment is expected to be more temporally stable whereas topical content e.g., natural disasters can change drastically over time and the works discussed in this section did not consider these issues in great detail.

[5] A program mapping biomedical text to concepts in the largest thesaurus in the biomedical domain [5]

2.5 Preferences and Traits

Learning user preferences will help us to sense what do users want e.g., what do they prefer to buy, what topics do they care about more. Social media provides sources of data for various topics, e.g., people reveal their preferences online, which can be mined.

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 [31]. The second case is when there are multiple users and multiple items. This scenario is often called collaborative filtering [37]. 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.

2.5.1 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 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.

2.5.2 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[43] 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[6]) shown with computation of Pearson's correlation[69]. Methods used by [43] and [69] 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. [43] used Facebook likes as the only feature, while Quercia et al. [69] used more extensive features including user's profile information, number of followers, and number of followees. Understanding users traits will bring insights on how to approach learning social sensors for various targets, and is a beginning step toward achieving superior topical social media sensors.

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 [91, 92], numeric real-values showing the preferences for each item [75], or binary values on whether the user would like an item or not [74]. They used different methodologies such as simple popularity methods [91], linear regression [91, 92, 74], simple classifiers (Naive Bayes, SVM, logistic regression) [91, 74], or collaborative filtering methods based on matrix factorization [75]. Zhang and Pennacchiotti [91, 92] use 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. [74] focuses on user interactions (type, modality, directionality) in addition

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

to user likes on Facebook. The study of users preferences on products could help topical social sensors by learning to return more personalized topical content for various users.

Political Preference Prediction Research on political preferences includes predicting political orientation [29, 30], classifying stances on political debates concerning topics of health care, gay rights, gun rights, ... [76, 81], 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. [30] applies collaborative filtering based on probabilistic matrix factorization.
- The non-collaborative works either use simple data mining and statistical approaches [29], homophily measure between users and their followers/followees using similarity metrics [1], or classification methods [76, 81]. To apply these methods, researchers extracted features including sentiment features [76, 81], and structure-based features (network of users on following each other) [29, 1].

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. Knowing user's political preferences could be useful in presenting users with more personalized search results on topics that are related to politics such as, Social Issues, Human Caused Disasters, etc. Hence, we could take advantage of the results from the classification of stances on political matters to build more personalized topical social sensors in future work.

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 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, 87, 67]: 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 [87] employed social-based features in addition to tweet author-

based features. Different from the other two works, Xu and Yang [87] performed the analysis from the perspective of individual users. Petrovic et al. [67] 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** [34, 25]: Works in this category focus on finding important tweets by analyzing propagation of tweets through retweeting. Feng and Wang [25] 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. [34] 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 [25] 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 2.1.

Despite the fact that all of these methods recommend tweets and take advantage of useful features such as tweet-based features and social based features, they and recommendation methods in general are focused on predicting tweets that correlate with the preferences of a specific user or that are directly related to specific content. Rather the focus with learning topical social sensors is to learn to predict for general topics (independent of a users profile) in a way that generalizes beyond existing labeled topical content to novel future topical content.

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 2.1: List of features used in retweet prediction

2.5.3 Summary

In this section we presented different methodologies and applications for preference detection and re-tweet recommendation in social media. Social sensors for preference detection are important since they provide us with more features on users preferences that could later be helpful for learning topical social sensors. These works raise a number of useful features such as user's traits, retweet probability of a tweet, social-based features that could be used in general social sensors, what remains is conducting a comprehensive evaluation of these features in general topic learning setting to see how well these features perform for detection of topical content.

2.6 Conclusion

Through review of the literature we showed that social media can be used as a sensor to detect latent phenomena. Existing works in the literature successfully detected or predicted events, sentiments, and preferences of users. The weaknesses of all the reviewed works could be summarized in a few points:

- Existing literature on trending topic detection is intrinsically unsupervised and not intended to detect targeted topics, which is in contrast to supervised learning of topical

social sensors for generic topics.

- Existing literature on physical event detection has the potential of providing high precision event detectors, however the existing literature is highly specific to the target event and does not easily generalize to learn arbitrary event-based or topic based social sensors as provided in this work.

- Existing literature on social sentiment and opinion mining discusses two sets of methdologies: (1) ad-hoc lexicon-based methods which are not extendable to learning general topical social sensors, and (2) using supervised classification methods with leveraging interesting features such as complex sentiment of social media posts. These features could be a useful to consider in learning topical social sensors, however evaluating the importance of the features for topicality remains as an open area for future research.

- Existing literature on preference learning could help social sensors in sensing what do users like on various matters like products, politics, etc. This in turn could be useful for learning topical social sensors to capture this information in providing better or more personalized topical content for users.

- Existing literature on tracking general topics provided a good basis for this work, however there are many fine-grained differences between our feature extraction and learning framework to the mentioned works. Such as the way we split our hashtags and tweets to train, validation and test sets to ensure generalization to future unseen content. We also provide a comprehensive feature analysis showing what/why features are useful for topical content detection while the literature has little to provide on this matter.

In the next chapter, we present our methodology for learning general topical social sensors.

Chapter 3: Learning Topical Social Sensors

The main focus of this thesis is retrieving high-precision content on generic topics where small amount of labeled data in form of hashtags is available from user. The question is how to learn over the space of millions of features to predict topical content with high precision that matches users information needs and will generalize to future unseen content. This chapter addresses resolving these issues in learning general topical social media sensors.

3.1 Problem Setup

In this section, we reduce the problem of learning topical content from large space of features and small set of examples provided by the user to the following setting that will match standard supervised learning paradigm.

Here, the problem statement is that the user has an information need for high-precision topical content from Twitter. The first step for the user is that he/she must provide labeled data to represent this information need for use in a targeted supervised learning setting. We assume that for each topic, the user will provide us with a set of hashtags. For example for the topic of *Natural Disaster*, the user will give us {#earthquake,#flood,#prayforthephilippines, ...}. The goal is, given this topic and related hashtags, return a ranked list of tweets to the user that are highly relevant to the topic and match users information needs; meaning instead of returning a set of tweets that only match ”disaster” or ”natural”, we realize the actual information need behind the searched topic and present all tweets matching this need. To this end, we need a methodology that learns from the set of hashtags provided by users on how to pick up sensors (i.e., useful terms, mentions, hashtags, locations, and users) and weight them to ensure picking new, unseen topical hashtags in future tweets. The following discussion intends to answer how to develop such methodology.

Our objective in learning social sensors is to train an automatic system for ranking documents by their topical relevance. Formally, given an arbitrary document d and a set of topic classes

$C = \{c_1, \ldots, c_K\}$, we wish to train a scoring function $f{:}d \rightarrow \mathbb{R}$ over a set of training documents $D = \{d_1, \ldots, d_N\}$. Considering M number of features extracted from dataset, each document $d_i \in D$ has a boolean feature vector $(d_i^1, \ldots, d_i^M) \in \{0,1\}^M$ and boolean label $d_i^c \in \{0,1\}$ indicating whether the document d_i is topical (1) or not (0). We define the set of positively occurring features for a document d_i as $D_i^+ = \{d_i^j | d_i^j = 1\}_{j=1\ldots M}$ and note that D_i^+ may include features for the content of d_i (e.g., terms, hashtags) as well as its meta-data (e.g., author, location).

There are two catches that make our training setting somewhat non-standard and which underlie subtle but critical contributions in this work:

1. Manually labeling documents is time-consuming so we need a way to label a large number of tweets with minimal user curation effort; *We achieve this by using hashtags as topical proxies.*

2. We need to train our social sensor on known topical content, but tune it on novel topical validation content that ensures the tuning achieves optimal generalization; *We achieve this by excising training content from our validation data so that our scoring function hyperparameter tuning ensures generalization.*

We next explain these key innovations in detail.

A critical bottleneck for learning targeted topical social sensors is to achieve sufficient supervised content labeling. With data requirements often in the thousands of labels to ensure effective learning and generalization over a large candidate feature space (as found in social media), manual labeling is simply too time-consuming for many users and crowdsourced labels are both costly and prone to misinterpretation of users' information needs. Fortuitously, hashtags have emerged in recent years as a pervasive topical proxy on social media sites — hashtags originated on IRC chat, were adopted later (and perhaps most famously) on Twitter, and now appear on other social media platforms such as Instagram, Tumblr, and Facebook. Hence as a simple enabling insight that serves as a catalyst for effective topical social sensor learning, for each topic class $c \in C$, we leverage a (small) set of user-curated topical hashtags H^c to efficiently provide a large number of supervised topic labels for social media content. Next we will provide the formal procedure for labeling data with H^c and training.

With the data labeling bottleneck resolved, we proceed to train supervised classification and

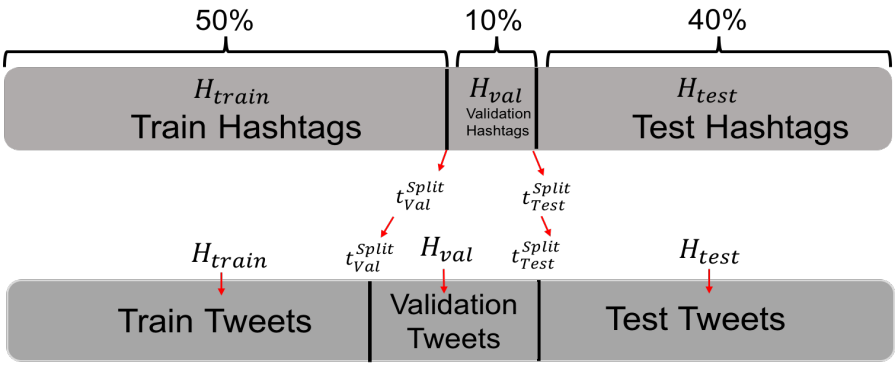


Figure 3.1: The method for temporally splitting hashtags and tweets to train, validation, and test sets

ranking methods to learn topical content from a large feature space (e.g., for Twitter, this feature space includes terms, hashtags, mentions, authors and their locations). The training process includes the following two steps:

1. **Temporally split train and validation using H^c:** As usual for machine learning methods, we divide our training data into train and validation sets — the latter for hyperparameter tuning to control overfitting and ensure generalization to unseen data. As a critical insight for topical generalization where we view identification of previously unseen hashtags as a proxy for topical generalization, we do not simply split our data temporally into train and test sets as usually done. Instead, we split H^c into two disjoint sets H^c_{train} and H^c_{val} according to a time stamp t_{split} and the first usage time stamp t^*_h of hashtags $h \in H^c$. This procedure is shown visually in Fig 3.1. Formally, we define the following:

$$H^c_{\text{train}} = \{h | h \in H^c \wedge t^*_h < t_{\text{split}}\},$$
$$H^c_{\text{val}} = \{h | h \in H^c \wedge t^*_h \geq t_{\text{split}}\}.$$

   Once we have split our hashtags into training and validation sets according to t_{split}, we next proceed to temporally split our training documents D into a training set D^c_{train} and a validation set D^c_{val} for topic c based on the posting time stamp t_{d_i} of each document d_i as

follows:

$$D^c_{\text{train}} = \{d_i | d_i \in D \wedge t_{d_i} < t_{\text{split}}\},$$
$$D^c_{\text{val}} = \{d_i | d_i \in D \wedge t_{d_i} \geq t_{\text{split}}\}.$$

Then for each set of *train* and *val* tweets, we use the respective hashtag sets H^c_{train} and H^c_{val} for labeling each $d^c_i \in D^c_{\text{train}}$:

$$d^c_i = \begin{cases} 1 : \exists_{h \in H^c_{\text{train}}} \; h \in D^+_i \\ 0 : \text{otherwise} \end{cases}.$$

and similarly for each $d^c_i \in D^c_{\text{val}}$:

$$d^c_i = \begin{cases} 1 : \exists_{h \in H^c_{\text{val}}} \; h \in D^+_i \\ 0 : \text{otherwise} \end{cases}.$$

The critical insight here is that we not only divide the train and validation temporally, but we divide the hashtag labels temporally and label the validation data with an entirely disjoint set of topical labels from the training data. The purpose behind this training and validation data split and labeling is to ensure that learning hyperparameters are tuned so as to prevent overfitting and maximize generalization to unseen topical content (i.e., new hashtags).

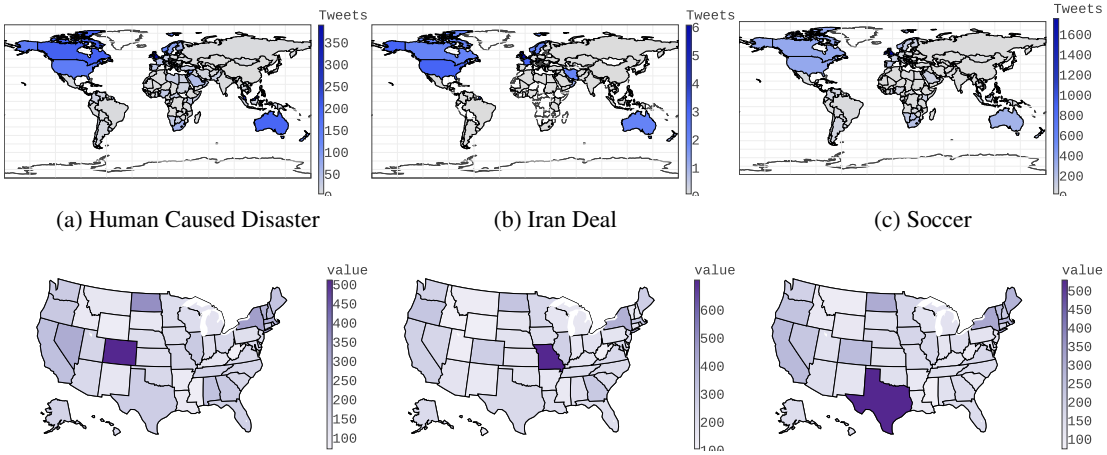
2. **Training and hyper-parameter tuning:** Once D^c_{train} and D^c_{val} have been constructed, we proceed to train our scoring function f on D^c_{train} and select hyperparameters to optimize Average Precision (AP) on D^c_{val}. Once the optimal f is found for D_{val}, we return it as our final learned topical scoring function for topic t.

Having defined our topical social sensor learning paradigm, it now remains to empirically evaluate this methodology in a social media setting, which we describe next.

3.2 Data Description

This section provides details of the Twitter testbed for topical social sensor learning that we evaluate in this thesis. We crawl Twitter data using Twitter Streaming API for two 24 months spanning 2013 and 2014. The total number of tweets collected is 829,026,458. In the context of Twitter, we consider five feature types for each tweet. Each tweet has a *From* feature (i.e., the person who tweeted it), a possible *Location* (i.e., a string provided as meta-data), and a time stamp when it was posted. A tweet can also contain one or more of the following:

- *Hashtag*: a topical keyword specified using the # sign.
- *Mention*: a Twitter username reference using the @ sign.
- *Term*: any non-hashtag and non-mention unigrams.



(a) Human Caused Disaster (b) Iran Deal (c) Soccer

(d) Health Epidemics (e) Social Issues (f) Space

Figure 3.2: Per capita tweet frequency across different international and U.S. locations. The Middle East and Malaysia stand out for Human Caused Disaster (MH370 incident), Iran and Europe for nuclear negotiations on "Iran deal", soccer for some (English-speaking) countries where it is popular. and on U.S. states, Colorado for health epidemics (both whooping cough and pneumonic plague), Missouri stands out for social issues (#blacklivesmatter in St. Louis), and Texas stands out for space due to NASA's presence

We provide more detailed statistics about each feature in Table 3.1. For example, there are over 11 million unique hashtags, the most frequent unique hashtag occurred in over 1.6 million

tweets, a hashtag has been used on average by 10.08 unique users, and authors (*From* users) have used a median value of 2 unique hashtags.

#Unique Features

From	Hashtag	Mention	Location	Term
95,547,198	11,183,410	411,341,569	58,601	20,234,728

Feature Usage in #Tweets

Feature	Max	Avg	Median	Max entity
From	10,196	8.67	2	running_status
Hashtag	1,653,159	13.91	1	#retweet
Mention	6,291	1.26	1	null
Location	10,848,224	9,562.34	130	london
Term	241,896,559	492.37	1	rt

Feature Usage by #Users

Hashtag	592,363	10.08	1	#retweet
Mention	26,293	5.44	1	dimensionist
Location	739,120	641.5	2	london
Term	1,799,385	6,616.65	1	rt

Feature Using #Hashtags

From	18,167	2	0	daily_astrodata

Table 3.1: Feature Statistics of our $829,026,458$ tweet corpus.

Fig. 3.2 shows per capita tweet frequency across different international and U.S. locations for different topics. While English speaking countries dominate English tweets, we see that the Middle East and Malaysia additionally stand out for the topic of Human Caused Disaster (MH370 incident), Iran and Europe for nuclear negotiations the "Iran deal", and soccer for some (English-speaking) countries where it is popular. For U.S. states, we see that Colorado stands out for health epidemics (both whooping cough and pneumonic plague), Missouri stands out for social issues (#blacklivesmatter in St. Louis), and Texas stands out for space due to NASA's presence

there.

With Twitter dataset explained, now we move to discussion of our proposed methodology for learning to rank high-value topical content in Twitter.

3.3 Proposed Approach

With the formal definition of learning topical social sensors provided in Sec. 3.1 and the overview of our data in Chapter. 4, we proceed to outline our experimental methodology on our Twitter corpus.

3.3.1 Dataset preparation

We manually curated a broad thematic range of 10 topics shown in the top row of Table 3.2 by annotating hashtag sets H^t for each topic $t \in T$. We used 4 independent annotators to query the Twitter search API to identify candidate hashtags for each topic, requiring an inner-annotator agreement of 3 annotators to permit a hashtag to be assigned to a topic set. Per topic, hashtags were split into train and test sets according to their first usage time stamp roughly according to a 3/5 to 2/5 proportion. The train set was further temporally subdivided into train and validation hashtag sets according to a 5/6 to 1/6 proportion. We show a variety of statistics and five sample hashtags per topic in Table 3.2. Here we can see that different topics had varying prevalence in the data with *Soccer* being the most tweeted topic and *IranDeal* being the least tweeted according to our curated hashtags.

3.3.2 Feature Extraction

As noted in Chapter. 4, positively occurring features D_i^+ in our d_i may include *From*, *Mention*, *Location*, *Term*, and *Hashtag* features. Because we have a total of $538,365,507$ unique features in our Twitter corpus, it is critical to pare this down to a size amenable for efficient learning and robust to overfitting. To this end, we thresholded all features according to the frequencies listed in Table 3.3. The rationale in our thresholding was initially that all features should have

	Tennis	Space	Soccer	IranDeal	HumanDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT
#TrainHashtags	58	98	126	12	49	28	31	31	52	29
#TestHashtags	36	63	81	5	29	16	19	19	33	17
#TopicalTweets	55,053	239,719	860,389	8,762	408,304	163,890	230,058	230,058	210,217	282,527
Sample Hashtags	#usopenchampion	#asteroids	#worldcup	#irandeal	#gazaunderattack	#robinwilliams	#policebrutality	#earthquake	#ebola	#loveislove
	#novakdjokovic	#astronauts	#lovesoccer	#iranfreedom	#childrenofsyria	#ripmandela	#michaelbrown	#storm	#virus	#gaypride
	#wimbledon	#satellite	#fifa	#irantalk	#iraqwar	#ripjoanrivers	#justice4all	#tsunami	#vaccine	#uniteblue
	#womenstennis	#spacecraft	#realmadrid	#rouhani	#bombthreat	#mandela	#freetheweed	#abfloods	#chickenpox	#homo
	#tennisnews	#telescope	#beckham	#nuclearpower	#isis	#paulwalker	#newnjgunlaw	#hurricanekatrina	#theplague	#gaymarriage

Table 3.2: Test/Train Hashtag samples and statistics.

	Threshold	#Unique Values
From	159	361,789
Hashtag	159	184,702
Mention	159	244,478
Location	50	57,767
Term	50	317,846
Features (CF)	-	1,166,582

Table 3.3: Cutoff threshold and corresponding number of unique values of candidate features *CF* for learning.

the same frequency cutoff in order to achieve roughly 1 million features. However, in initial experimentation, we found that a high threshold pruned a large number of informative terms and locations. To this end, we lowered the threshold for terms and locations noting that even at these adjusted thresholds, we still have more authors than terms. We also removed common English stopwords which further reduced the unique term count. Overall, we end up with $1,166,582$ candidate features (*CF*) for learning social sensors.

3.3.3 Supervised Learning Algorithms

With our labeled training and validation datasets defined in Sec. 3.1 and our candidate feature set *CF* defined previously, we proceed to apply different probabilistic classification and ranking algorithms to generate a score function f for learning social sensors as defined in Sec. 3.1. In this paper, we experiment with the following four state-of-the-art classification and ranking methods:

1. **Logistic Regression** using LibLinear [23]
2. **Bernoulli Naïve Bayes** [54]
3. **Rocchio** [50]
   (a centroid-based classifier)
4. **RankSVM** [45]

As outlined in Sec 3.1, tuning of hyperparameters on a validation dataset is critical. In our experiments, we tune the following hyperparameters:

- *Logistic Regression*: L_2 regularization constant C is tuned for $C \in \{1E-12, 1E-11, ..., 1E+11, 1E+12\}$.
- *Naïve Bayes*: Dirichlet prior α is tuned for $\alpha \in \{1E-20, 1E-15, 1E-8, 1E-3, 1E-1, 1\}$.
- *All Classfiers*: The number of top features M selected based on their Mutual Information is tuned for $M \in \{1E2, 1E3, 1E4, 1E5, 1166582 \text{ (all features)}\}$.

We remark that many algorithms such as Naive Bayes and Rocchio performed better with feature selection and hence we used feature selection for all algorithms (where it is possible to select all features). Hyperparameter tuning is done via exhaustive grid search and using the Average Precision (AP) to select the best scoring function f on the validation data. Once found, f can be applied to any tweet d_i to provide a score $f(d_i)$ used to rank tweets in the test data.

3.4 Performance Analysis

We now proceed to evaluate the performance of each of the four aforementioned supervised learning algorithms for the task of learning social sensors. We note that parts of the analysis was conducted using Apache Spark on Amazon Web Services to handle the large amount of data. For evaluation, two main tasks are considered. (1) The correct classification of each tweet (topical or not), (2) Evaluation of the ranking of the returned tweets for each topic. Once a scoring function is trained via each method, we use it to rank tweets and then compute the following ranking metrics on the resulting ranked list:

- **AP:** Average precision over the ranked list; the mean over all topics provides mean AP (MAP).
- **P@k:** Precision at k for $k \in \{10, 100, 1000\}$.

While P@10 may be a more standard retrieval metric for tasks such as ad-hoc web search, we remark that the short length of tweets relative to web documents makes it more plausible to look at a much larger number of tweets, hence the reason for also evaluating P@100 and P@1000.

Table 3.4 evaluates these metrics for each topic. *Logistic Regression* is the best performing method on average except for $P@10$. We conjecture the reason for this is that *Naïve Bayes* tends to select fewer features for training, which allows it to achieve higher precision over the top of the ranked list but which causes it suffer slightly more lower down the list due to having fewer features and lower recall. These results suggest that in general both *Logistic Regression* and *Naïve Bayes* make for effective topical social sensor learners with *Naïve Bayes* being a good choice in terms of its efficiency compared to it's overall performance.

To provide more insight into the general performance of our learning topical social sensor framework, we provide the top five tweets for each topic returned by *Logistic Regression* in Table 3.5 and Table 3.6. We've annotated all tweets in this table with the following symbols:

- ✓: the tweet was topical according to our curated test hashtag set.
- ★: the tweet was determined to be topical through manual evaluation even though it did not contain a hashtag in our curated hashtag set (this corresponds to a false negative due to non-exhaustive labeling of the data).
- ✗: the tweet was not topical.

In general, we remark that our learning social sensor based on logistic regression performs even better than the quantitative results in Table 3.4 would indicate: many of the highly ranked tweets are false negatives — they are actually relevant. Furthermore, we remark that even though we use hashtags to label our training, validation, and testing data, our learning social sensor has highly (and correctly) ranked topical tweets that do not contain hashtags indicating encouraging generalization properties from a relatively small set of curated topical hashtags.

3.5 Discussions

In this chapter we presented the proposed framework and methodology for learning topical social sensors from Twitter. Here, we provided a supervised learning method which is novel in structure and application evidenced by the following highlights: (1) the use of a set of hashtags as topical proxies to label millions of tweets with minimal user effort, (2) the novel way of temporally splitting hashtags to use them for splitting and labelling train, validation, and test

datasets of tweets and later removing the tweets containing train hashtags from the evaluation set, and (3) The largely impressive results across 10 diverse topics in terms of average precision and precision@n.

In the next chapter, we provide a through analysis of our features explaining how features and their attributes correlate with topicality in Twitter.

		Tennis	Space	Soccer	IranDeal	HumanDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT	Mean
LR	**AP**	**0.918**	0.870	0.827	0.811	0.761	0.719	0.498	**0.338**	**0.329**	**0.165**	**0.623±0.19**
NB	**AP**	0.908	**0.897**	0.731	**0.824**	**0.785**	**0.748**	**0.623**	0.267	0.178	0.092	0.605±0.22
Rocchio	**AP**	0.690	0.221	**0.899**	0.584	0.481	0.253	0.393	0.210	0.255	0.089	0.407±0.18
RankSVM	**AP**	0.702	0.840	0.674	0.586	0.603	0.469	0.370	0.248	0.136	0.082	0.471±0.18
LR	**P@10**	**1.000**	0.000	0.200	0.700	**0.600**	0.000	0.100	0.200	0.300	**0.500**	0.360±0.24
NB	**P@10**	**1.000**	**0.900**	0.700	0.600	**0.600**	**0.700**	**1.000**	0.100	0.400	0.100	**0.610±0.23**
Rocchio	**P@10**	0.800	0.000	**1.000**	**0.900**	0.000	0.000	0.000	**0.500**	**0.500**	0.100	0.380±0.29
RankSVM	**P@10**	**1.000**	0.800	0.600	0.800	0.400	0.300	0.000	0.100	0.000	0.200	0.420±0.26
LR	**P@100**	0.950	0.580	0.650	0.870	0.620	0.490	0.640	**0.690**	**0.790**	**0.210**	**0.649±0.15**
NB	**P@100**	**0.980**	**0.850**	0.600	**0.880**	0.750	**0.860**	**0.730**	0.230	0.090	0.190	0.616±0.23
Rocchio	**P@100**	**0.980**	0.000	**1.000**	0.690	0.170	0.000	0.280	0.170	0.680	0.120	0.409±0.28
RankSVM	**P@100**	0.730	0.720	0.310	0.700	**0.880**	0.440	0.480	0.340	0.020	0.100	0.472±0.20
LR	**P@1000**	**0.963**	**0.954**	0.816	**0.218**	0.899	0.833	**0.215**	0.192	**0.343**	**0.071**	**0.550±0.26**
NB	**P@1000**	0.954	**0.954**	0.716	**0.218**	**0.904**	**0.881**	**0.215**	**0.195**	0.141	0.060	0.524±0.28
Rocchio	**P@1000**	0.604	0.000	**0.925**	**0.218**	0.359	0.000	**0.215**	0.167	0.144	0.065	0.270±0.21
RankSVM	**P@1000**	0.799	0.922	0.764	**0.218**	0.525	0.547	**0.215**	0.173	0.154	0.064	0.438±0.22

Table 3.4: Performance of topical social sensor learning algorithms across metrics and topics with the mean performance over all topics shown in the right column. The best performance per metric is shown in bold.

Tennis
✓rt @espntennis: shock city. darcis drops rafa in straight sets. first time nadal loses in first rd of a. major...
✓@ESPNTennis: Shock city. Darcis drops Rafa in straight sets. First time Nadal loses in first rd of a...
✓@ESPNTennis: Djokovic ousts the last American man standing @Wimbledon, beating Reynolds 7-6...
✓Nadal's a legend. After 3 years; Definitely He's gonna be the best of all the time. Unbelievable perf...
✓@calvy70 @ESPNTennis @Wimbledon I see, thanks for the info and enjoy #Wimbledon2014
Soccer
✗rt @tomm_dogg: #thingstodobeforeearthends spend all my money.
★@mancityonlineco nice performance
★rt @indykaila: podolski: "let's see what happens in the winter. the fact is that i'm not happy with it, th...
★rt @indykaila: wenger: "i don't believe match-fixing is a problem in england." #afc
✗@indykaila you never got back to me about tennis this week
HumanDisaster
✓rt @baselsyrian: there'ye been peaceful people in #homs not terrorists! #assad,enemy of #humanity...
✓what a helpless father, he can do nothing under #assad's siege!#speakup4syrianchildren http://t.co/vg...
★exclusive: us formally requested #un investigation; russia pressured #assad to no avail;chain of evidence...
★#save_aleppo from #assadwarcrimes#save_aleppo from #civilians -targeted shelling of #assad regime...
✓rt @canine_rights: why does the #un allow this to continue? rt@tintin1957 help raise awareness of the...
SocialIssues
★the us doesn't actually borrow is the thing. i believe in a creationist theory of the us dollar @usanationdebt...
★rt @2anow: according to @njsenatepres women's rights do not include this poor nj mother's right to defend...
★rt @2anow: confiscation ? how many carry permits are in the senate and assembly? give us ours or turn ...
★rt @2anow: vote with your wallet against #guncontrolforest city enterprises does not support the #2a http...
★@2anow @momsdemand @jstines3 they dont have a plan for that,which is why they should never be allow...
Epidemics
✓rt @who: fourteen of the susp. & conf. ebola cases in #conakry, #guinea, are health care workers, of...
✗@who who can afford also been cover in government health insurance [with universal health coverage]
✓#ebolaoutbreak this health crisis..unparalleled in modern times, @who dir. aylward - requires $1 billion ...
✗rt @medsin: @who are conducting a survey on the social determinants of health in medical teaching. fill...
✗augmentation vertigineuse de 57,4% en 1 an des actes islamophobes en france, dit le collectif contre l'is...

Table 3.5: Top tweets for each topic from *Logistic Regression* method results, marked with ✗as irrelevant, ✓as relevant and labeled as topical, and ★as relevant but labeled as non-topical

Space
✗rt @jaredleto: rt @30secondstomars: icymi: mars performing a cover of @rihanna's #stay on australia's @trip...
✗voting mars @30secondstomars @jaredleto @shannonleto @tomofromearth xobest group http://t.co/dls...
✗rt @jaredleto_com: show everyone how much you are proud of @30secondstomars !#mtvhottest 30 seconds to..
✗rt @30secondstomars: missed the big news? mars touring with @linkinpark + special guests @afi this sum...
✗rt @30secondstomars: to the right,to the left,we will fightto the death.go #intothewildonvyrt with mars, starting...
IranDeal
✓rt @iran_policy: @vidalquadras:@isjcommittee has investigated 10 major subjects of irans controversial #nuc...
✓rt @iran_policy: @vidalquadras:@isjcommittee has investigated 10 major subjects of irans controversial #nuc...
✗rt @negarmortazavi: thank you @hassanrouhani for retweeting. let's hope for a day when no iranian fears retur...
✗rt @iran_policy: iran: details of savage attack on political prisoners in evin prison http://t.co/xdzuakqdiv #iran...
✓rt @iran_policy: chairman ros-lehtinen speaking on us commitment 2 protect camp liberty residents. #iranhr...
CelebrityDeath
★rt @sawubona_chris: today is my birthday & also the day my hero @nelsonmandela has died. lets never...
★rt @nelsonmandela: death is something inevitable.when a man has done what he considers to be his duty to...
★rt @nelsonmandela: la muerte es algo inevitable.cuando un hombre ha hecho lo que considera que es su...
✗#jacques #kallis: a phenomenal cricketing giant of all time - #cricket #history #southafrica http://t.co/ms5p...
✗@sudesh1304 south africa has the most beautiful babies....so diverse,so unique...so god!! lol #durban #southa...
NaturalDisaster
✗us execution in #oklahoma : not cruel and unusual? maybe just barbaric, inhumane and reminiscent of the...
✗#haiti #politics - the haiti-dominican crisis - i agree with how martelly is handling the situation: i totally... http...
★rt @soilhaiti: a new reforestation effort in #haiti. local compost, anyone? http://t.co/xpad0rqbjk @richardbran...
✗mes cousins jamais ns hantent les nuits de duvalier #haiti #duvalier
✓tony burgener of @swisssolidarity says you can't compare the disaster response in #haiti with the response to...
LGBT
★rt @jackmcoldcuts: @lunaticrex @fingersmalloy @toddkincannon @theanonliberal anthony kennedy just...
✗@toddkincannon your personal account, your interest. separate from your business.
✗why would you report someone as spam if he is not spam? @illygirlbrea @toddkincannon
✗rt @t3h_arch3r: @toddkincannon thanks for your tl having the female realbrother. between them is 600 lbs....
✗@toddkincannon who us dick trickle.

Table 3.6: Top tweets for each topic from *Logistic Regression* method results, marked with ✗as irrelevant, ✓as relevant and labeled as topical, and ★as relevant but labeled as non-topical

Chapter 4: Feature Analysis

In this chapter, we analyze the informativeness of our defined features in Chapter 4 and the effect of their attributes on learning targeted topical social sensors. To this end, our goal in this section is to answer the following questions:

1. What are the best features for learning social sensors and do they differ by topic?
2. For each feature type, do any attributes correlate with importance?

To answer these questions, we use Mutual Information (MI) [51] as our primary metric for feature evaluation. Mutual Information is defined as a general method for measuring the amount of information one random variable contains about another random variable. Mutual Information has been highly successful for feature selection , hence it serves as a measure of feature utility for the topic classification task [32]. In order to calculate the amount of information that each feature $j \in \{\textit{From}, \textit{Hashtag}, \textit{Mention}, \textit{Term}, \textit{Location}\}$ provides w.r.t. each topic label $t \in \{NaturalDisaster, Epidemics, ...\}$, Mutual Information is formally defined as

$$I(j,t) = \sum_{t \in \{\text{true,false}\}} \sum_{j \in \{\text{true,false}\}} p(j,t) \log \left(\frac{p(j,t)}{p(j)p(t)} \right),$$

where higher values for this metric indicate more informative features for the specified topic.

4.1 Feature Importance

In order to answer the first question regarding the best features for learning social sensors, we provide the mean Mutual Information values for each feature across different topics in Fig. 4.1. The last column in Fig. 4.1 shows the average of the mean Mutual Information for each feature type. From analysis of Table 4.1, we can make a set of observations:

- The *Term* and *Location* features are the most informative features on average.

Topics/Top10	NaturalDisaster	Epidemics	IranDeal	SocialIssues	LBGT
From	earthquake_wo	changedecopine	mazandara	nsingerdebtpaid	eph4_15
From	earthalerts	drdaveanddee	hhadi119	debtadvisoruk	mgdauber
From	seelites	joinmentornetwk	140iran	debt_protect	stevendickinson
From	globalfloodnews	followebola	setarehgan	negativeequityf	lileensvf1
From	gcmcdrought	localnursejobs	akhgarshabaneh	dolphin_ls	truckerbooman
Hashtag	earthquake	health	iran	ferguson	tcot
Hashtag	haiyan	uniteblue	irantalks	mikebrown	p2
Hashtag	storm	ebola	rouhani	ericgarner	pjnet
Hashtag	tornado	healthcare	iranian	blacklivesmatter	uniteblue
Hashtag	prayforthephilippines	depression	no2rouhani	fergusondecision	teaparty
Location	philippines	usa	tehran	st.louis	usa
Location	ca	ncusa	u.s.a	mo	bordentown
Location	india	garlandtx	nederland	usa	newjersey
Location	newdelhi	oh-sandiego	iran	dc	sweethomealabama!
Location	newzealand	washington	globalcitizen	washington	aurora
Mention	oxfamgb	foxtramedia	4freedominiran	deray	jjauthor
Mention	weatherchannel	obi_obadike	iran_policy	natedrug	2anow
Mention	redcross	who	hassanrouhani	antoniofrench	govchristie
Mention	twcbreaking	obadike1	un	bipartisanism	a5h0ka
Mention	abc7	c25kfree	statedept	theanonmessage	barackobama
Term	philippines	health	iran	police	obama
Term	donate	ebola	regime	protesters	gun
Term	typhoon	acrx	nuclear	officer	rights
Term	affected	medical	iranian	protest	america
Term	relief	virus	resistance	cops	gop

Topics/Top10	HumanDisaster	CelebrityDeath	Space	Tennis	Soccer
From	ydumozyf	nmandelaquotes	daily_astrodata	tracktennisnews	losangelessrh
From	syriatweeten	boiknox	freesolarleads	tennis_result	shoetale
From	tintin1957	jacanews	houston__jobs	i_roger_federer	sport__agent
From	sirajsol	ewnreporter	star_wars_gifts	tennislessonnow	books_you_want
From	rt3syria	paulretweet	lenautilus	kamranisbest	makeupbella
Hashtag	syria	rip	science	wimbledon	lfc
Hashtag	gaza	riprobinwilliams	starwars	usopen	worldcup
Hashtag	isis	ripcorymonteith	houston	tennis	arsenal
Hashtag	israel	mandela	sun	nadal	worldcup2014
Hashtag	mh370	nelsonmandela	sxsw	wimbledon2014	halamadrid
Location	malaysia	southafrica	germany	london	liverpool
Location	palestine	johannesburg	roodepoort	uk	manchester
Location	syria	capetown	houston	india	london
Location	israel	pretoria	austin	pakistan	nigeria
Location	london	durban	tx	islamabad	india
Mention	ifalasteen	nelsonmandela	bizarro_chile	wimbledon	lfc
Mention	revolutionsyria	realpaulwalker	nasa	usopen	arsenal
Mention	drbasselabuward	robinwilliams	j_ksen	andy_murray	realmadriden
Mention	mogaza	rememberrobin	jaredleto	serenawilliams	ussoccer
Mention	palestinianism	tweetlikegiris	30secondstomars	espntennis	mcfc
Term	israel	robin	cnblue	murray	madrid
Term	gaza	williams	movistar	tennis	goal
Term	israeli	nelson	enero	federer	cup
Term	killed	mandela	imperdible	djokovic	manchester
Term	children	cory	greet	nadal	match

Table 4.1: The top 5 features for each feature type and topic based on Mutual Information. We note that the *Term*s appear to be the most generic and generalizable features, and the top *Location*s are also highly relevant to most topics indicating the overall importance of these tweet features for identifying topical tweets.

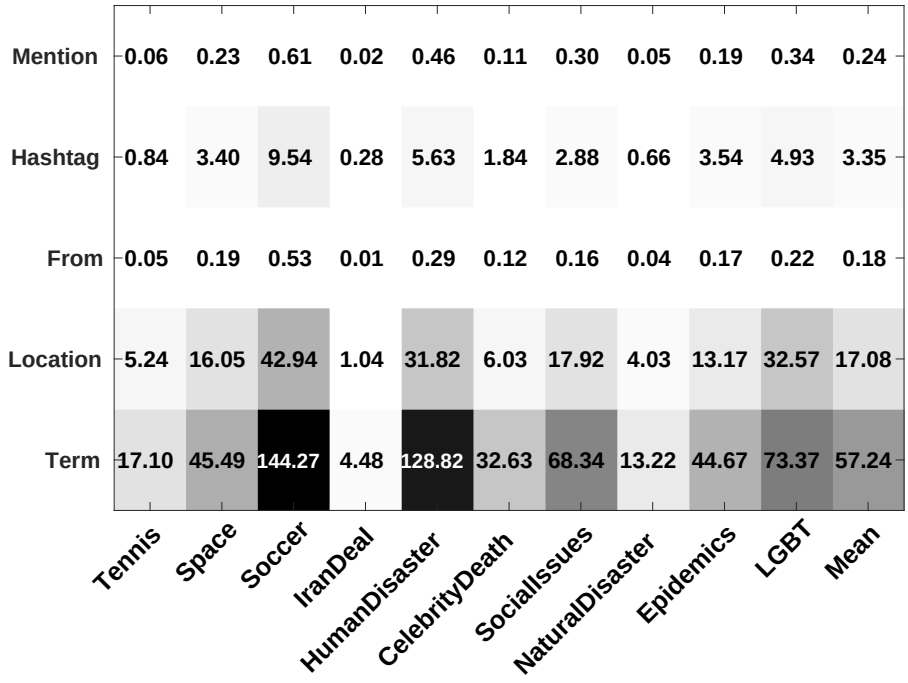


Figure 4.1: Matrix of mean Mutual Information values for different feature types vs. topics. The last column as average of mean values across all topics. All values should be multiplied by $1E+10$. We remark that the *Term* and *Location* features are the most informative features on average and the *Location* feature provides the most information regarding the topics of *HumanDisaster*, *LBGT*, and *Soccer* indicating that a lot of content in these topics is heavily localized)

- The *Location* feature provides the most information regarding the topics of *HumanDisaster*, *LBGT*, and *Soccer* indicating that a lot of content in these topics is heavily localized.
- Looking at the overall average values, the order of informativeness of feature types appears to be the following: *Term*, *Location*, *Hashtag*, *Mention*, *From*.

To further analyze the relationship between the informativeness of feature types and topics, we refer to the box plots of Fig. 4.4. Here we see the quartiles and outliers of the distribution rather than just the average of the MI values in order to ensure the mean MI values were not misleading our interpretations. Overall, however, the story is the same: *Term* and *Location* features dominate

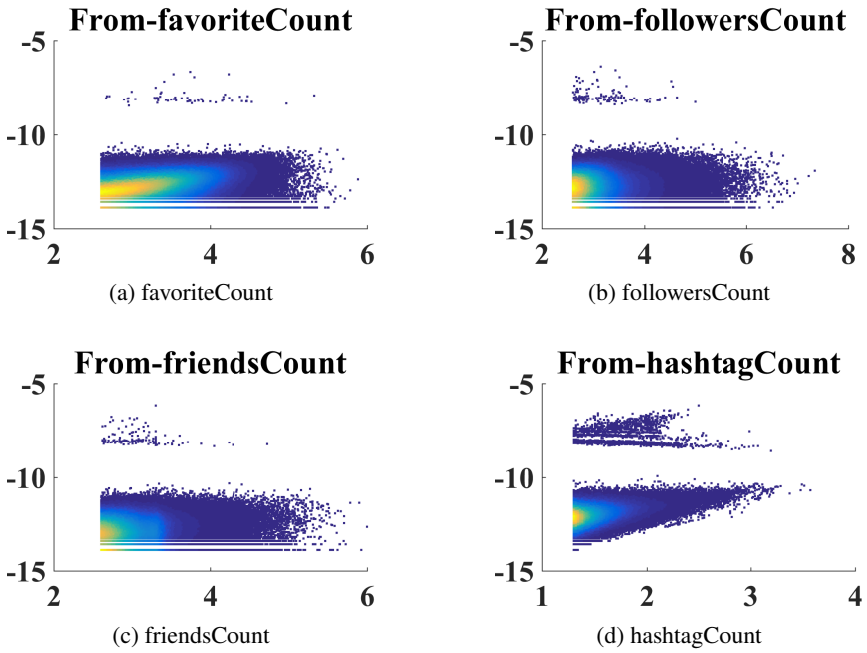
in terms of Mutual Information followed by the other less informative features. Furthermore, two observations are apparent: (1) *Term*s have more outliers indicating that the most useful individual features may be terms, and (2) the topic has little impact on which feature is most important indicating stability of feature type informativeness over topics.

As anecdotal evidence to inspect which features are most informative, we refer to Table 4.1, which displays the top five feature instances for each feature type and topic. Among many remarkable insights in this table, one thing we note are that the *Term*s appear to be the most generic (and hence most generalizable) features, providing strong intuition as to why these features figure so prominently in terms of their informativeness. The top *Location*s are also highly relevant to most topics indicating the overall importance of these tweet features for identifying topical tweets.

4.2 Attribute Importance

In order to answer the second question on whether any attributes correlate with importance for each feature, we provide two types of analysis. The first analysis shown in Fig. 4.3 analyzes the distributions of Mutual Information values for features when binned by the magnitude of various attributes of those features, outlined as follows:

- **From** vs.
  - *Favorite count:* # of tweets user has favorited.
  - *Followers count:* # of users who follow user.
  - *Friends count:* # of users followed by user.
  - *Hashtag count:* # of hashtags used by user.
  - *Tweet count:* # of tweets from user.
- **Hashtag** vs.
  - *Tweet count:* # of tweets using hashtag.
  - *User count:* # of users using hashtag.



(a) favoriteCount (b) followersCount

(c) friendsCount (d) hashtagCount

Figure 4.2: Density plots for the Mutual Information vs. frequency values of feature attributes. Plots (a-d) respectively show attributes {favoriteCount, followerCount, friendCount, hashtagCount} for the *From* feature. We remark that an interesting bimodalilty is clear from these plots. Our analysis showed that the the top mode feature occurs in at least one topical tweet whereas the bottom mode occurs in no topical tweets.

- **Location** vs. *User count:* # of users using location.
- **Mention** vs. *Tweet count:* # of tweets using mention.
- **Term** vs. *Tweet count:* # of tweets using term.

As we can see in the Violin plots of Fig. 4.3, the general pattern is that the greater the number of tweets, users, or hashtag count a feature has, the more informative the feature is in general. This pattern also exists to some extent on the attributes of the *From* feature, although the pattern is less visible in general and not clear (or very weak) for the follower or friend count. In general, the informativeness of a user appears to have little correlation with their follower or friend count.

Fig. 4.2 provides a further analysis by showing density plots of favorite count, follower count, friends count, and hashtag count attributes of the *From* feature. Here we see an interesting phenomenon that was not clear in the Violin plots: there is a very clear bimodality of the density. On further investigation it turns out that the top mode feature occurs in at least one topical tweet whereas the bottom mode occurs in no topical tweets. While the bottom mode features may serve as good indicators of non-topicality, the top mode are inherently more indicative of topicality, which justifies feature selection by mutual information.

4.3 Summary

In this chapter, we provided a comprehensive study on the features extracted from tweets and their attributes to evaluate their correlation with topicality. We draw a number of important insights from this analysis, the summary of them is as following:

- Computing average Mutual Information values for different feature types vs. topics shows that independent of topic, *Term*s and *Location*s are the most important features. Anecdotal examples for top 5 features of each topic also represent that *Term* features are the most generic and hence the most generalizable features in general. Further, we note that *Location* feature is more informative for a few topics compared to other ones which demonstrates that lots of content in topics such as *HumanDisaster*, *LBGT*, and *Soccer* are localized.
- Plotting the distribution of Mutual Information values for features vs. magnitude of their

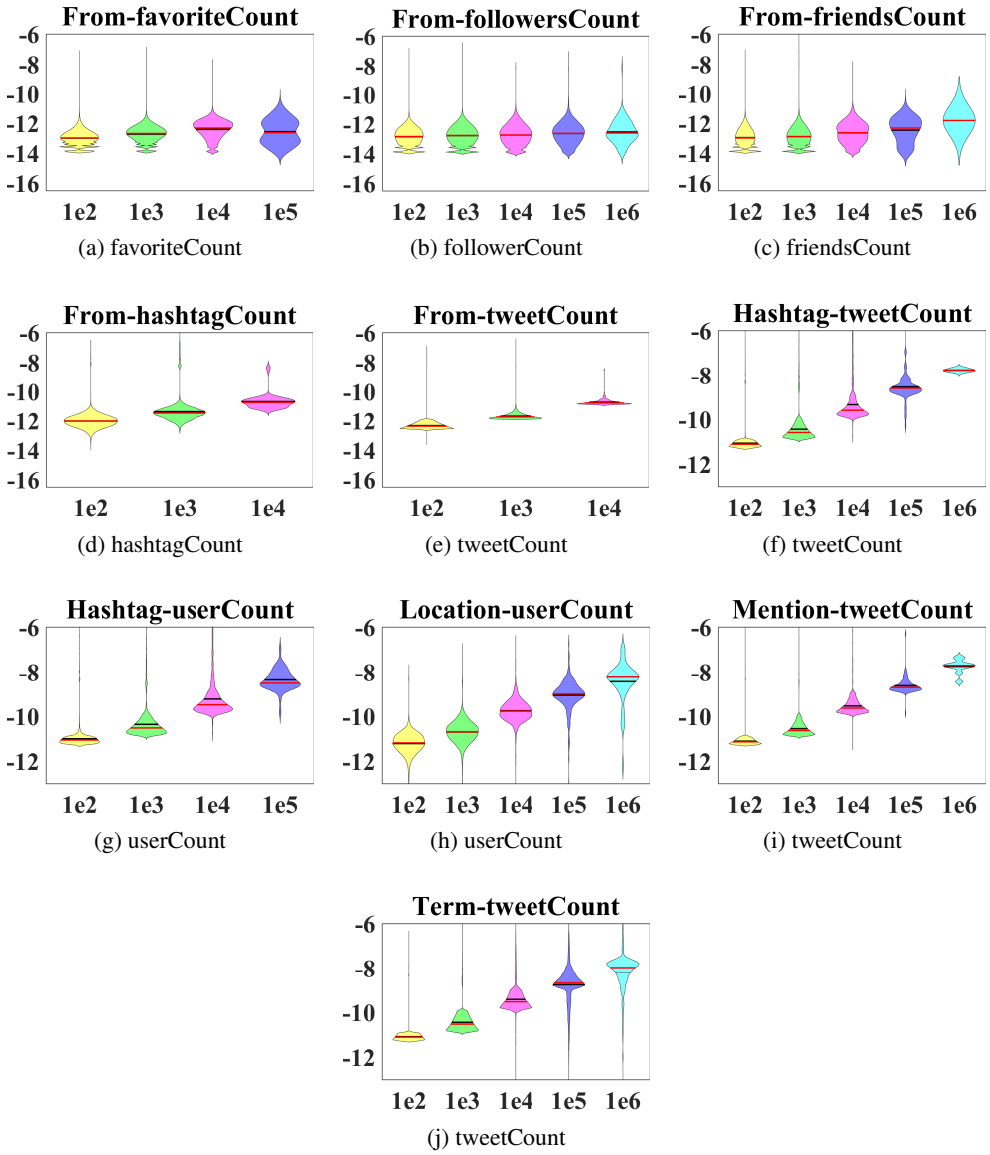


Figure 4.3: Violin plots for the distribution of Mutual Information values of different features as a function of their attributes. Plots (a-e) respectively show attributes {favoriteCount, followerCount, friendCount, hashtagCount, tweetCount} for *From* feature. We note that the the higher the number of tweets and hashtags, the more important a *From* feature is, however, the informativeness of a user appears to have little correlation with their follower or friend count. Plots (f-j) respectively show attributes tweetCount and userCount for *Hashtag*, userCount for *Location* feature, tweetCount for *Mention* and *Term* features. We remark that the general pattern for attributes of the features is that the greater the number of tweets, users or hashtag counts a feature has, the more informative it is.

attributes shows that the greater the number of tweets, users, or hashtags counts of a feature, the more important the feature is. Interestingly enough, number of followers and friends counts of a user does not demonstrate the same pattern and in general there is very little or no correlation between these attributes and a user's importance.

- Despite the fact that hashtags were used to label the data, interestingly it was one of the less important feature types. This shows that the proposed method of labeling training data does not create a bias to overfit to training hashtags and it is conjectured that this helps explaining the excellent generalization to data labeled with unseen test hashtags.

In the next chapter, we present the conclusion of the thesis and future works.

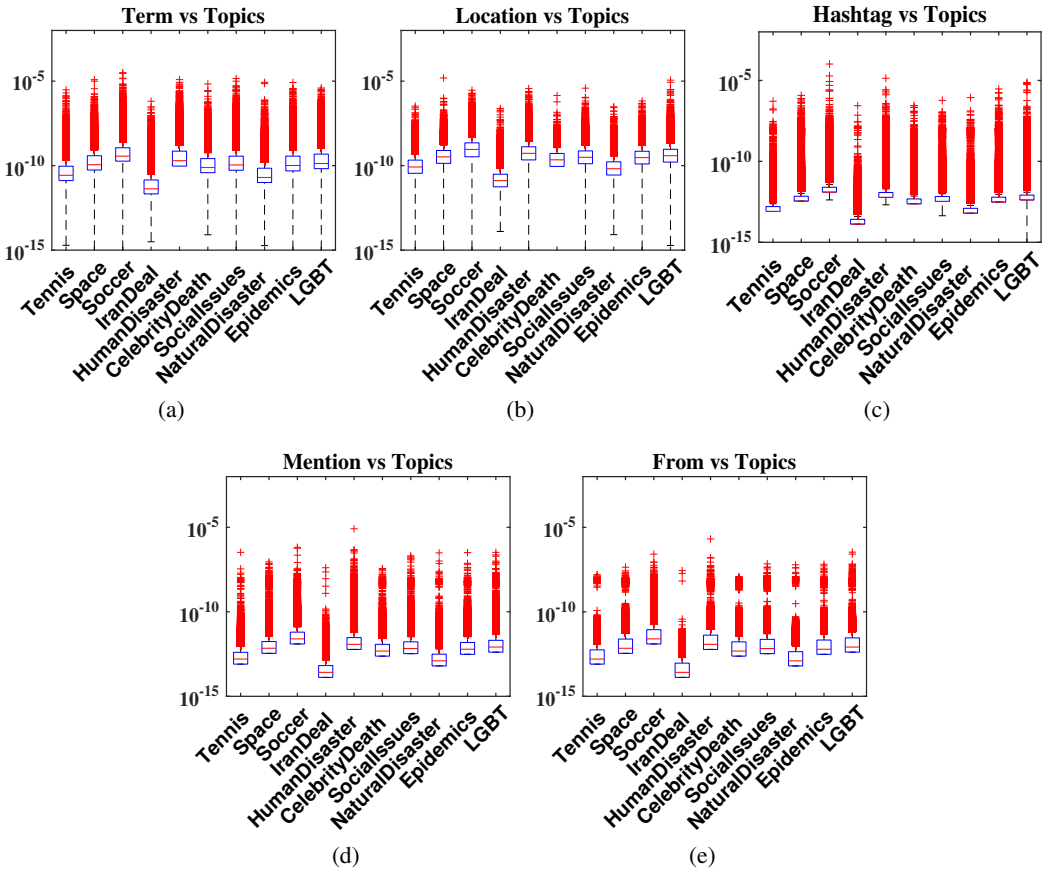


Figure 4.4: Box plots of Mutual Information values per feature type across topics. We remark that *Term*s have more outliers indicating that the most useful individual features may be terms, and the topic has little impact on which feature is most important indicating stability of feature type informativeness over topics.

Chapter 5: Conclusion

5.1 Summary of contributions

In this thesis, we aimed to address supervised learning of topical social media sensor with the purpose of detecting and ranking topical content from Twitter for general topics. In chapter 2, we discussed that the concept of social media as a sensor has been used in the literature and introduces the use of social media as sensors for detecting news, events, opinions, etc. However, our literature review presented a gap in these works showing how these methods fail either by not being targeted to any specific topic, such as trending topic detection, being too narrow and specific on a single topic, such as earthquake detection, or because of using very primitive methods for collecting their data such as methods for sentiment detection. Further, we discussed that this thesis builds on and extends a handful of works [47, 49, 88] that have been done on learning topical social sensors.

In this thesis, we made the following contributions:

- In Chapter 3, we proposed a novel supervised learning method for training social sensors on a dataset of 800 million tweets crawled from Twitter. The novelty of this method stems from (1) the way we label large quantities of data using a small set of hashtags as topical proxies for automatic data labeling, (2) the way we train our models where we temporally split hashtags and tweets into train, validation, and test sets to cover two years of data, and (3) the way we evaluate generalization of our methodology to future unseen topical content by measuring average precision and precision@n only on tweets that omitt training hashtags. We remark that two simple and efficiently trainable methods, logistic regression and Naïve Bayes, were capable of learning users information needs and generalize well to unseen future topical content (including content with no hashtags) in terms of their mean average precision (MAP) and Precision@n for a range of n.

- In Chapter 4, we provided a comprehensive study on the features and their attributes extracted from our Twitter dataset. The analysis involved two major parts: (1) Analysis of the

correlation of each feature with topicality, (2) Analysis on whether different attributes of each feature have correlations with topicality, for example whether having more followers would make a person to be a more important feature.

Our results suggest that:

- Learned social sensors generalize well to unseen future topical content and provide a novel paradigm for the extraction of high-value content from social media. We remark that out of four different methods, *Logistic Regression* and *Naïve Bayes* match users information needs and generalize well to future content in terms of their mean average precision (MAP) and Precision@n for a range of n. Furthermore, our results shows that we could learn generalizable sensors beyond train hashtags and beyond hashtags in general.
- Extensive analysis of features and feature attributes across different topics has revealed many useful insights among which the two important insights are:
  1. Largely independent of topic, simple terms are the most informative feature followed by location features.
  2. Interestingly, the number of unique hashtags and tweets by a user correlates more with their informativeness than their follower or friend count.

5.2 Future Work

This work uncovers some interesting areas for future work. In particular, future work should focus on exploring the following enhanced topical social sensor learning tasks.

1. **Beyond Boolean relevance**:

   This thesis used a Boolean relevance model. However in practice, topical relevance tends to be more graded – by degree of relevance, importance, temporal novelty. This brings out a need to figure out how to obtain graded relevance. Assuming lag-time of a topic is the time between when it has been mentioned for the first time and when we detect and return it, temporal novelty can be evaluated for hashtags as lag-time since first usage of the hashtag. To this end, we need to leverage regression methods that are well matched to the

data type (e.g., truncated, ordinal, etc.) as well as using ranking metrics for non-Boolean relevance.

2. **Learning Boolean queries**:

   The methodology explained in this thesis is based on the assumption that we have access to a large dataset of tweets either through accessing Twitter's firehose[1], or like our case by streaming a large dataset of tweets from Twitter. However, when we only have access to Twitter's search API, we need an approach that optimizes queries for Boolean retrieval oriented APIs while maintaining high precision and recall. To the best of our knowledge, no existing work has been applied to learning Boolean queries. Learning Boolean queries is a challenging task since it does not support learning different weights for features and not all features are equally important and thus weighting the features is critical in learning topical social sensors. To this end, we need to explore methods with hypothesis spaces that match Boolean queries better than log-linear classifiers, such as decision tree approaches. Two possible directions for learning Boolean queries could be:

   (a) Learn a single compact nested query: This approach would require us to learn one single query that is comprised from OR of a set of AND clauses for each topic that would query Twitter search API once for the topic. This could be achieved by learning Decision Trees or Gradient Boosted Regression Trees [26].

   (b) Merge the results of multiple queries: This approach will benefit from learning multiple different queries for a specific topic followed by merging and ranking the results. This approach could help in the sense that it will allow us to weight the results of each query and therefore to partially approximate weighting in the weighted feature learning methods such as logistic regression.

3. **Learning for long-term stability**:

   Some sensors may display a sharp decay in performance over time because features were very localized in time. For example, the feature *location:philippines* could be a good sensor for detection of *Natural Disasters* right after a flood happened in Philippines, whereas a feature like *term:donate* is a good generic sensor irrespective of what point in time it has

[1] https://dev.twitter.com/streaming/firehose

been used. This brings us to the question of how to learn temporally stable features.To this end, we need to formulate an evaluation metric to measure and rank features based on their long-term stability. The evaluation metric should be able to consider power of features for extraction of topical content in different time windows. We need to use this evaluation metric as an objective for training more stable topical social sensors.

Altogether, we believe this and future work will pave the way for a new class of social sensors that learn to identify broad themes of topical information with minimal user interaction and enhance the overall social media user experience.

Bibliography

[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: Detecting 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. ACM, 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] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

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

[25] 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.

[26] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

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

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] Ching-Pei Lee and Chih-Jen Lin. Large-scale linear RankSVM. *Neural Computing*, 26(4): 781–817, April 2014.

[46] 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.

[47] Jimmy Lin, Rion Snow, and William Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 422–429. ACM, 2011.

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

[49] Walid Magdy and Tamer Elsayed. Adaptive method for following dynamic topics on twitter. In *ICWSM*, 2014.

[50] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

[51] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

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

[53] 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.

[54] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *In AAAI-98 Workshop On Learning For Text Categorization*, pages 41–48. AAAI Press, 1998.

[55] 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.

[56] 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.

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

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

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

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

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological on-line 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.

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

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

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

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM, 2014.

[89] 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.

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

[91] 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.

[92] 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.

[93] 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.

APPENDICES

Appendix A: Supporting Theories

This chapter discusses a range of theories that support applications of using social media as a sensor presented in the literature review on Chapter 2.

A.0.1 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 [86], (2) affective or emotional state [62], or (3) the intended emotion the author wanted to convey [35]. (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 2.4.

A.0.1.1 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 [62], in a study of text polarity showed human's sentimental reaction to various words across eight dimensions, for example, three dimensions are the following:

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 A.1: Positive and negative side of 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 and negative side of each dimension are shown in table A.1 [33]:

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 [86]. In general, appraisal theory is an analysis of how a writer values people and things within the text that he/she produces [52]. 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 [62]'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 [57, 41] 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 [78]. 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 [78] 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

A.0.2 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.

A.0.2.1 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, 83]. 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 A.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 A.1.

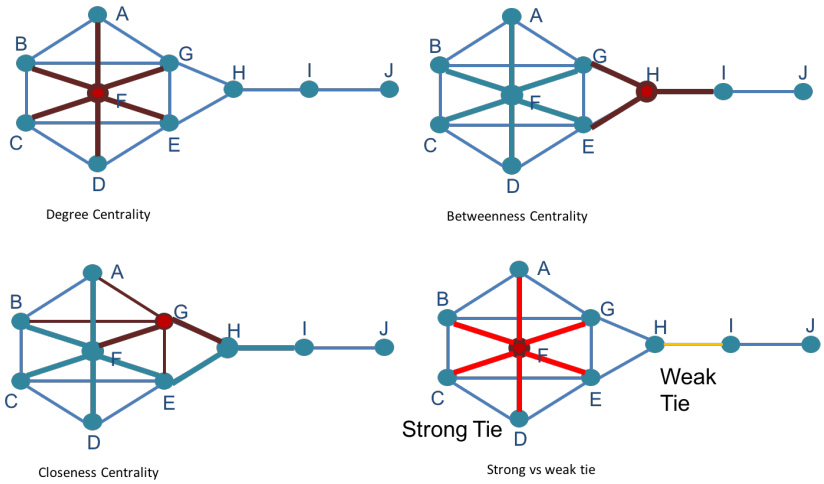


Figure A.1: Different centrality criteria and Weak vs. Strong ties

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[1], 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" 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

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

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 [83]. 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[2], 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.

[2] Albert-Laszlo Barabasi and Eric Bonabeau

Feld [24] 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 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 [24].

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 2.2.2.

A.0.2.2 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 [58] introduced an underlying contact networks model [58]; 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 [56]. 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 [58]
2. SIR: Susceptible-Infected-Removed [42]
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 [84]. This is of high importance in studying how news, video, and opinion become viral [46, 7].

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 [70]. 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, 36].

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.

Appendix B: Notations

Meaning	**Notation**
Topics	$C = \{c_1, \ldots, c_K\}$
Documents	$D = \{d_1, \ldots, d_N\}$
Set of hashtags of topic class c	H^c
Set of train hashtags of topic c	H^c_{train}
Set of validation hashtags of topic c	H^c_{val}
Time of the first usage of hashtag $h \in H^c$	t^*_h
Posting time of tweet d_i	t_{d_i}
Set of training tweets of topic c	D^c_{train}
Set of validation tweets of topic c	D^c_{val}
Positive occurring features for document d_i	$D_i^+ = \{d_i^j \vert d_i^j = 1\}_{j=1\ldots M}$
Boolean label for document d_i and topic c	$d_i^c \in \{0, 1\}$

Table B.1: Notations used and their meaning