TwittDict: Extracting Social Oriented Keyphrase Semantics from Twitter

Suppawong Tuarob[†], Wanghuan Chu[‡], Dong Chen[§], and Conrad S Tucker[♯]
[†]Faculty of Information and Communication Technology, Mahidol University, Thailand
[‡] Department of Statistics, [§]Information Sciences and Technology,
[♯]Industrial and Manufacturing Engineering, Pennsylvania State University, USA
suppawong.tuarob@gmail.com, {wxc228,duc196,ctucker4}@psu.edu

Abstract

Social media not only carries information that is up-to-date, but also bears the wisdom of the crowd. In social media, new words are developed everyday, including slangs, combinations of existing terms, entity names, etc. These terms are initially used in small communities, which can later grow popular and become new standards. The ability to early recognize the existence and understand the meanings of these terms can prove to be crucial, especially to emergence detection applications. We present an ongoing research work that investigates the use of topical analysis to extract semantic of terms in social media. In particular, the proposed method extracts semantically related words associated with a target word from a corpus of tweets. We provide preliminary, anecdotal results comprising the semantic extraction of five different keywords.

1 Introduction

Multiple applications built upon social media data have emerged and recently gained attention from a wide range of research fields. For example, public surveillance systems have shown success in employing Twitter data to detect the emergence of diseases (Tuarob et al., 2013b; Tuarob et al., 2014), emergency needs during natural disasters (Caragea et al., 2011), and even changes in product trends (Tuarob and Tucker, 2015c; Tuarob and Tucker, 2015a). Regardless of such appealing applications, tremendous challenges exist in employing traditional natural language processing techniques to handle social media data. Most of the issues with social media involve language creativity and noise, such as non-standard terms or symbolic expressions, caused by the users. Languages in social media evolve rapidly as the users have the freedom to express their opinions in colloquial, everyday languages. Some social media services such as Twitter limit the length of each message, that even challenge the users to express their complete thoughts in a compressed manner, resulting in creativity that would be considered noise by most traditional NLP techniques. This language evolution can be classified into two categories: grammatical alteration and word distortion. Grammatical alteration involves incomplete sentences (e.g. `'Dance Practice All Day Hit Up The iPhone4 (:'`), omitting words or part of words (e.g. `'[Does] Anyone have suggestions for [an] iPhone 4 mic?'`), and developing new terms (e.g. `'I totally fricken agree!'`). Word distortion involves modifying existing terms to deviate from the original meanings or to encode a phrase into a single word, such as `looooooove` (much love) and `lol` (laugh out loud). Besides the language evolution, noise is also considered a norm in social media. The sources of such noise include the use of symbolic representations (e.g. `':)'`) and typographical errors (both by intention and unintention). Both language evolution and noises produce non-standard terms, words not defined in a standard dictionary. Moreover, non-standard terms may refer to proper nouns, or entity names, e.g. *Xbox*, *Microsoft*, and *Peking*. These non-standard terms pose challenges to existing semantic interpretation techniques, especially those dependent on dictionary look-up of terms.

Text normalization techniques such as those utilizing noisy channel models (Cook and Stevenson, 2009; Xue et al., 2011) rely on the assumption that a non-standard term has its equivalent standard form (e.g. *love* $\Rightarrow$ *looooove*). With such an assumption, the algorithms aim to reverse the transformation process and seek the original

Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, pages 25–31,
Beijing, China, July 30, 2015. ©2015 Association for Computational Linguistics

form of a non-standard term. These algorithms, however, would fail if a term is newly developed and does not have a counterpart standard form (for example, `'swine flu'`, `'linsanity'`, `'Tweeps'`, etc.).

In particular, we present *TwittDict*, a model for semantic exploration of unknown terms in social media. Specifically, the method first identifies different topics discussed in the social media corpus. It assumes that a given term is associated with one or more topics, which then allows the mapping between such a term with relevant topically represented terms. Though multiple works have shown success on semantic annotation of unknown terms, these works target the domain of traditional documents where noise and language evolution are not taken into account. A preliminary case study that uses Twitter data to extract semantically relevant terms from a set of chosen five target terms is presented.

2 Background and Related Work

Use of social media, such as collaboratively edited knowledge databases (Wikipedia[1]), blogs and microblogs (Biyani et al., 2014), content communities (YouTube[2]), and social networking sites (Facebook[3]) (Kaplan and Haenlein, 2009), has grown at a prodigious rate. According to Nielsen's report[4], the total amount of time spent by the U.S. population on social media in 2012 was 520.1 billion minutes, a 21% increase from the previous year. This results in the creation and diffusion of a huge amount of information on social media everyday, including news, knowledge, opinions, and emotions. Different groups use social media for different reasons. For instance, companies can use social media to gather customer feedback and conduct market research and reputation management. Governmental organizations can spread news and gather public opinions. Meanwhile, the wealth of information on social media contributes to the collective wisdom and can be used to predict real-world outcomes such as stock prices (Bollen et al., 2011), flu trends (Lampos et al., 2010), and product sales (Tuarob and Tucker, 2013). To realize the potential of social media, the first step is to select relevant information, which requires an understanding of language evolution on social media. One aspect of such evolution is the creation and use of new terms aiming at describing timely events or new social phenomenon. Many of these terms are too new to be indexed by standard dictionaries or Wikipedia, and the results returned by popular search engines like Google[5] can be obscure and unstructured. Therefore, we seek to use social community knowledge to extract term semantics which provide better understanding on the language evolution.

2.0.1 Semantic Discovery of Terms

Weischedel et al. (1993) had success in employing probabilistic models to discover unknown terms and annotate them with parts of speech. Daniel et al. (1999) proposed a named entity recognition (NER) algorithm which categorizes a proper noun into one of the 3 predefined categories: Location, Person, and Organization. Besides Daniel et al's work, other NER algorithms such as (Chieu and Ng, 2002) achieved similar goals. These solutions rely on the assumption that a proper noun must fall into one of the predefined categories, while it is ubiquitous to see new categories of terms emerge from social media. Moreover, these algorithms require the data to adhere to standard English grammar. This requirement is hardly satisfied in social media. Fellbaum described *Wordnet*[6] a lexical database for English vocabulary that provides a set of synonyms (synset) for a given word. However, such database is constructed manually and only contains standard dictionary words, while our solution is fully automatic and can be applied to standard and non-standard terms that appear in social media.

2.0.2 Quantifying Unknown Terms in Social Media Data

Dealing with non-standard terms can be cumbersome. Dictionary-based approaches tend to fail when facing such unknown terms since they basically do not exist and cannot be looked up. Cook and Stevenson (2009) identified 10 different ways in which a term can be distorted in mobile text messaging. They proposed a noisy channel unsupervised model to translate a non-standard term into its standard version. Xue et al. (2011) proposed a similar channel-based model to translate a non-standard term into its standard form in the

[1] http://www.wikipedia.org/
[2] https://www.youtube.com/
[3] https://www.facebook.com/
[4] http://www.nielsensocial.com/
[5] https://www.google.com/
[6] http://wordnet.princeton.edu/

Twitter domain. These algorithms assume that an unknown term can be mapped one-to-one to its standard form. Unfortunately, the presence of newly generated terms naturally found in social media violate such an assumption, simply because these terms are newly developed and hence do not have their standard forms. These newly developed terms include social slangs, trending words, and names of entities.

Lund and Burgess (1996) attempted to explore the semantic of terms by generating the term occurrence network. A term is annotated with its highly related terms based on the distances in the network. Though their algorithm treats a document as a bag of words (hence does not rely on sentence structures), the algorithm produces meaningful results when the data is high-dimensional and dense. Such properties result in a strong and meaningful co-occurrence relationship. However, each message in social media is usually represented with a short text, resulting in high-dimensional but sparse data. Consequently, such data sparsity would impede the co-occurrence relationship.

3 Methodology

Topic models (Blei and Lafferty, 2009) are powerful tools to study latent patterns in text. The semantic of an unknown word is highly related to the topics associated with the text that contains it. Moreover, identified topics can be considered as representatives of the semantic. While one document might only have a limited number of topics associated with it, the collection of a large amount of documents containing the unknown term can provide more thorough and comprehensive understanding. Therefore, topic models can be applied to extract the semantics of unknown terms with large enough collection of documents. Social media such as Twitter usually adopts the use of newly developed terms at a very fast rate. Social media users tweet about topics related to the unknown terms based on their subjective understanding. Different tweets may present different meanings towards a single term. While a single tweet lacks the information to provide the full semantics of the term, a collection of all the tweets containing the term would give a much larger and clearer picture of the semantics. Therefore, topic models can be applied on social media to extract word semantics in terms of collective wisdom and social knowledge.

In this study, we choose the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model topical variation due to its flexibility and richness in the results. We use Twitter data as a case study, hence the name *TwittDict* is devised. Note that our algorithm can also be applied to other social media such as Facebook and Google+, as long as the medium of communication is in textual forms and community structures exist. In this section, we first briefly review our problem and introduce the LDA model, and then discuss how we filter the related tweets and how we apply the LDA to extract word semantics.

3.1 Problem Definition

Given a query word, TwittDict outputs a list of related words associated with it. The output words are ranked according to their relevance to the input term. Specifically, let $D_t = \{d_1, d_2, ..., d_n\}$ be the set of tweets, where each tweet $d_i \in D_t$ is a bag of words, W the vocabulary extracted from D_t, and w_t the query word. The proposed algorithm aims to output a ranked list of K words which are semantically relevant to w_t. For example, given a word `'Linsanity'`, the proposed algorithm would return a ranked list of semantically relevant words {`basketball, player, insanity, scholarship`} (with $K = 4$) as the output.

3.2 Latent Dirichlet Allocation

In text mining, the Latent Dirichlet Allocation (LDA) is a generative model that allows a document to be represented by a mixture of topics. The basic intuition of LDA for topic modeling is that an author has a set of topics in mind when writing a document. A topic is defined as a distribution of terms. The author then chooses a set of terms from the topics to compose the document. With such assumption, the whole document can be represented using a mixture of different topics. LDA serves as a means to trace back the topics in the author's mind before the document is written. Mathematically, the LDA model is described as follows:

$$P(w_i|d) = \sum_{j=1}^{|Z|} P(w_i|z_i = j) \cdot P(z_i = j|d). \quad (1)$$

$P(w_i|d)$ is the probability of term w_i being in document d. z_i is the latent (hidden) topic. $|Z|$ is the

number of all topics, which needs to be predetermined. $P(w_i|z_i = j)$ is the probability of term w_i being in topic j. $P(z_i = j|d)$ is the probability of picking a term from topic j in the document d.

Essentially, the aim of LDA model is to find $P(z|d)$, the topic distribution of document d, with each topic described by the distribution over all terms $P(w|z)$.

After the topics are modeled, we can assign a distribution of topics to a given document using a technique called *inference*. A document then can be represented by a vector of numbers, each of which represents the probability of the document belonging to a topic:

$$\text{Infer}(d, Z) = \langle z_1, z_2, ..., z_Q \rangle; |Z| = Q,$$

where Z is a set of topics, d is a document, and z_i is a probability of the document d falling into topic i. We use the Latent Dirichlet Allocation algorithm to generate topics in our model since it allows a topic to be represented by a distribution of terms, enabling the method to propagate the relevance from the target term to the underlying terms that compose the relevant topics.

3.3 Data Preprocessing

Twitter data is collected using the Twitter API. The textual information in each tweet is first lowercased, then usernames, stopwords, punctuations, numbers, and URLs are removed. While using the wealth of information on Twitter to understand an unknown term, the first step is to filter in tweets that are related to such a term. The most intuitive collection consists of all the tweets that contain the target word and treats each single tweet as a document, which we call the basis setting. However, there are some special characteristics of Twitter messages that we want to consider for modifications and improvements. First, there is limited information within each tweet because of the 140-character restriction, and the average length of tweets is even smaller. This is quite different from the traditional uses of the LDA where input documents are rich (e.g., research articles, newspaper, etc), and hence generated topics are quite intuitive and meaningful. Second, other information in tweets such as *retweet* (`RT`), *reply* (`@username`) and *hashtag* (`#`) exist, which can be used more appropriately instead of just being deleted or treated as a plain word. To overcome the drawbacks and make better use of Twitter features, we consider improving the basis setting by expanding the collection of tweets using *reply* and *hashtag*. *Reply* refers to those tweets that start with `@username` and comment on other tweets. For the tweet that contains the unknown term, its *reply* tweets make comments on the same or other related topics. Although these tweets might not contain the target word, it is reasonable to assume that they should be in similar semantic as the original tweet thus providing additional information. Therefore, we will expand the collection of tweets by combining all the *reply* tweets to the original one which contains the target term. *Hashtag* can also be used to find related tweets. People use the hashtag symbol # before a relevant keyword or phrase without space in the tweets to facilitate automatic categorization and search. These hashtags can be viewed as topical markers, serving as indications to the context or the core idea of the tweet. Tweets with the same hashtag share similar topics. Therefore, we use hashtags in the basis tweet to find all the other tweets that have at least one of these hashtags, which also enriches the information in the collection.

3.4 Retrieving Related Words

Mathematically, given a target document corpus $D_t = \langle d_1, d_2, ..., d_n \rangle$ (as described in Section 3.3), vocabulary $W = \langle w_1, w_2, ..., w_m \rangle$, and target word w_t, our algorithm outputs a ranked list $W_K^* = \langle w_1, w_2, ..., w_K \rangle$, where $w_i \in W$, of K words relevant to w_t.

Our algorithm comprises two main steps:

1. $P(w|w_t, W, D_t)$, the likelihood probability of the word w being relevant to the target word w_t, is computed for each $w \in W$.

2. Return top K words ranked by the likelihood probability.

In general, $P(w|w_t, W, D_t)$ is computed by weighted averaging of the posterior probability of $P(w|Z)$ across the documents in D, where Z is the set of topics:

$$P(w|w_t, W, D_t) = \sum_{z \in Z} P(z|D_t) \cdot P(w|z), \quad (2)$$

where $P(w|z)$ is the posterior probability of the word w being in topic z, computed in Equation 1. $P(z|D_t)$ serves as the weight of the topic z,

computed by averaging out the topic probability $P(z|d)$ across all documents in D_t:

$$P(z|D_t) = \frac{1}{|D_t|} \sum_{d \in D_t} P(z|d), \quad (3)$$

where $P(z|d)$ is computed based on Equation 1. Hence:

$$P(w|w_t, W, D_t) = \frac{1}{|D_t|} \sum_{z \in Z} \sum_{d \in D_t} P(z|d) \cdot P(w|z) \quad (4)$$

4 Evaluation

TwittDict is evaluated against the baseline which utilizes a variant of word co-occurrence to retrieve relevant keywords. Church et al. had success on using the mutual information to extract semantic related terms (1990). Furthermore, Tuarob and Tucker had used the word co-occurrence network to explicate implicit semantics in product related tweets (2015b). Here, the word co-occurrence network is constructed from the tweet corpus. The co-occurrence network is an undirected graph where each node is a distinct word, and each edge weight represents the frequency of co-occurrence. The edge weights can be used directly to compute $P(x, y)$, where x and y are co-occurred words. Given a target word w_t, a corpus of tweets T, and vocabulary $W = \langle w_1, w_2, ..., w_m \rangle$, the baseline algorithm outputs a ranked list $W_K^B = \langle w_1, w_2, ..., w_K \rangle$, where $w_i \in W$, of K words relevant to w_t. The algorithm assigns a co-occurrence based score to each word, and rank them by such a score. In this work, we experiment with three variations of co-occurrence based scores: Mutual Information (MI), Co-Frequency (CoF), and Co-Frequency Inverse Document Frequency (CoF-IDF):

$$Score_{MI}(w_t, w) = \log_2 \frac{P(w_t, w)}{P(w_t) \cdot P(w)} \quad (5)$$

$$Score_{CoF}(w_t, w) = P(w_t, w) \quad (6)$$

$$Score_{CoF-IDF}(w_t, w) = P(w_t, w) \cdot IDF(w, T) \quad (7)$$

5 Preliminary Case Study

We experiment our methodology with Twitter data and a set of manually selected words. Twitter data is used due to its ubiquitousness and public availability. Note that, our methodology can expand to other types of social media such as Facebook and Google+ if the data is available.

5.1 Twitter Data

Twitter is a microblog service that allows its users to send and read text messages of up to 140 characters, known as tweets. The Twitter dataset used in this research study comprises roughly 700 million tweets in the United States during the period of 19 months, from March 2011 to September 2012.

5.2 Anecdotal Results

A set of five target words ($Obama$, $Pandora$, $Xbox$, $Glee$, and $Zombie$) are used to test our proposed algorithm against one of the baseline with Co-frequency scores. TwittDict employs the LDA implementation in Mallet[7], with 100 topics and runs for 1,000 iterations using Gibb's Sampling. Due to the limitation on the computational time, TwittDict currently only models topics from a tweet corpus collected in March 2011. For the baseline, we first index the whole tweet corpus using Apache Lucene[8], then use the same library to compute word frequency. Table 1 lists the results.

From the preliminary results, TwittDict is able to extract highly meaningful words related to the target words, while the baseline contain a mixture of both related and generally spurious words. Note that, TwittDict only uses one month's worth (5.26%) of the available Twitter data, as opposed to the baseline which uses the whole collection of tweets. It is our belief that, with more Twitter data, TwittDict could even provide a wider variety and higher in semantics of lexicons.

6 Conclusions and Future Works

By leveraging natural language processing techniques and specific features in social media, we have described our ongoing development of *TwittDict*, a system to identify the social-oriented semantic meaning of unknown words. Such a system could prove to be useful as a building block for emergence detection systems where early recognition of new terms/concepts is crucial. We illustrated through anecdotal results using Twitter data to identify semantic meanings of five terms, that our method is not only achieving promising results, but also urging us to explore further into improving our methods along with conducting rigorous user and automatic evaluations such as (Tuarob et al., 2013a; Tuarob et al.,

[7] http://mallet.cs.umass.edu/

[8] http://lucene.apache.org/

Table 1: Preliminary results of 5 test words using both the baseline (CoF scores) and TwittDict.

Word	Obama		Pandora		Xbox		Glee		Zombie	
/Rank	Co-Freq	TwittDict	Co-Freq	TwittDict	Co-Freq	TwittDict	Co-Freq	TwittDict	Co-Freq	TwittDict
1	president	president	flow	station	live	live	watching	watching	apocalypse	apocalypse
2	vote	libya	station	radio	play	play	love	tonight	feel	lol
3	michelle	people	listening	listening	playing	kinect	watch	episode	lol	www
4	romney	war	radio	lol	got	lol	tonight	watch	day	movie
5	barack	bush	love	music	lol	playing	episode	love	dead	today
6	lol	barack	point	playing	time	game	season	song	mode	feel
7	don	don	tonight	song	need	games	project	time	sleep	movies
8	america	news	commercials	love	game	time	lol	good	walking	love
9	love	pres	playing	time	add	black	time	show	time	time
10	speech	time	lol	songs	don	back	omg	lol	today	back
11	fuck	gop	song	good	kinect	don	cast	songs	movie	zombies
12	got	white	time	shit	buy	buy	wait	don	night	band
13	dnc	america	listen	listen	games	good	song	night	zombies	dead
14	voting	oil	songs	day	fuck	day	good	cast	don	day
15	people	world	shit	today	shit	ops	don	amazing	love	mode
16	good	tcot	night	play	wanna	gamertag	amazing	omg	good	horror
17	campaign	japan	music	work	day	follow	night	week	shit	good
18	years	administration	jamming	flow	love	win	version	version	rob	house
19	win	house	got	tonight	controller	controller	excited	awesome	walkingdead	plays
20	osama	gas	sleep	night	haha	love	week	music	need	atomic

2015). There is plenty of room to improve *TwittDict*. In the current case study, we only used Twitter data during Mar 2011. This specific period of time may bring about bias towards the result. To avoid such bias, we need to test data in different times and geographical regions. This will shed light on how meanings of a term evolve temporally and spatially. When we were conducting the small case study, we noticed that the results were highly dependent on the time period, as Twitter users usually tweet about the current social phenomena. This change reflects the evolvement of social events and community knowledge. We are considering giving users the freedom to specify the time period during which a term is defined. Furthermore, we would explore methods for user evaluations. We would recruit human participants to give feedback about their experience. Real user experience is of great value for us to see whether and how community knowledge from social media truly helps them to better understand the unknown, emerging concepts. Finally, we would like to compare our method against well-established baseline such as (Turney et al., 2010) and (Mikolov et al., 2013).

References

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Mach. Learn.*, 34(1-3):211–231, February.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities.

David M Blei and J Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications*, 10:71.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

C. Caragea, N. McNeese, A. Jaiswal, G. Traylor, H.W. Kim, P. Mitra, D. Wu, A.H. Tapia, L. Giles, B.J. Jansen, et al. 2011. Classifying text messages for the haiti earthquake. In *ISCRAM '11*.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. CALC '09, pages 71–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas M. Kaplan and Michael Haenlein. 2009. The fairyland of second life: Virtual social worlds and how to use them. *Business Horizons*, 52(6):563 – 572.

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. 2010. Flu detector-tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. Springer.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Suppawong Tuarob and Conrad S Tucker. 2013. Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. In *ASME IDETC/CIE '13*.

Suppawong Tuarob and Conrad S Tucker. 2015a. Automated discovery of lead users and latent product features by mining large scale social media networks. *Journal of Mechanical Design*.

Suppawong Tuarob and Conrad S Tucker. 2015b. A product feature inference model for mining implicit customer preferences within large scale social media networks. In *ASME IDETC/CIE '15*.

Suppawong Tuarob and Conrad S Tucker. 2015c. Quantifying product favorability and extracting notable product features using large scale social media data. *Journal of Computing and Information Science in Engineering*.

Suppawong Tuarob, Line C Pouchard, and C Lee Giles. 2013a. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. JCDL '13, pages 239–248.

Suppawong Tuarob, Conrad S Tucker, Marcel Salathe, and Nilam Ram. 2013b. Discovering health-related knowledge in social media using ensembles of heterogeneous features. In *CIKM '13*, pages 1685–1690. ACM.

Suppawong Tuarob, Conrad S Tucker, Marcel Salathe, and Nilam Ram. 2014. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of biomedical informatics*, 49:255–268.

Suppawong Tuarob, Line C Pouchard, Prasenjit Mitra, and C Lee Giles. 2015. A generalized topic modeling approach for automatic document annotation. *International Journal on Digital Libraries*, 16(2):111–128.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Ralph Weischedel, Richard Schwartz, Jeff Palmucci, Marie Meteer, and Lance Ramshaw. 1993. Coping with ambiguity and unknown words through probabilistic models. *Comput. Linguist.*, 19(2):361–382, June.

Zhenzhen Xue, Dawei Yin, and Brian D Davison. 2011. Normalizing microtext. In *Proceedings of the AAAI Workshop on Analyzing Microtext*, pages 74–79.