Analyzing and Retrieving Illicit Drug-Related Posts from Social Media

Tao Ding, Arpita Roy, Zhiyuan Chen, Qian Zhu, Shimei Pan
The Department of Information Systems
University of Maryland, Baltimore County
Baltimore, MD 21250
{taoding01,arpita2,zhchen,qianzhu,shimei}@umbc.edu

***Abstract*—Illicit drug use is a serious problem around the world. Social media has increasingly become an important tool for analyzing drug use patterns and monitoring emerging drug abuse trends. Accurately retrieving illicit drug-related social media posts is an important step in this research. Frequently, hashtags are used to identify and retrieve posts on a specific topic. However hashtags are highly ambiguous. Posts with the same hashtags are not always on the same topic. Moreover, hashtags are evolving, especially those related to illicit drugs. New street names are introduced constantly to avoid detection. In this paper, we employ topic modeling to disambiguate hashtags and track the changes of hashtags using semantic word embedding. Our preliminary evaluation shows the promise of these methods.**

I. Introduction

Illicit drug use has become a severe problem which has a big impact on many areas of our society: crime, health care, child welfare and more. According to the results of the annual National Survey on Drug Use and Health (NSDUH) in 2014, an estimated 27.0 million people aged 12 or older used an illicit drug in the past 30 days[1]. This percentage in 2014 was higher than those in every year from 2002(8.3%) through 2013(9.4%). Since many factors may contribute to an individual's drug use such as biological and social factors, to study the effects of various social factors on drug use, many existing researches rely on data collected via surveys or interviews [1][2][3]. Since survey or interview-based data collection is expensive and time-consuming, it's hard to do it in large scale without proper incentives.

With the advent of social networks, millions of people share their status updates every day. Among these posts, many mention drug use. Due to its scale and instantaneous nature, social media has increasingly become an important information resource for analyzing drug use patterns and detecting emerging drug abuse trends. Specifically, social media data are

1) **easily accessible**: most popular social media sites such as Facebook, Twitter, Instagram provide APIs to allow easy access to their public contents.
2) **large scale**: these social platforms have hundreds of millions of users. For example, the Twitter community includes 120 million active users worldwide. They post more than 5.5 million tweets every day. With 1.55 billion active users, if Facebook is a nation, it is the largest on earth.
3) **in real time**: social media data can be streamed in real-time thus a good source of timely information, which is critical for monitoring emerging drug use trends.

To facilitate information retrieval, most social media platforms provide search APIs to support hashtag-based retrieval. However, hashtag-based retrieval is often noisy and ambiguous. For example, #ChinaGirl is the street name for the drug "fentanyl". As you can imagine, many posts matching this hashtag will not be related to "fentanyl". In addition, if you use the hashtag "#fentanyl" to search relevant social media posts, not all of them are related to illicit drug use since fentanyl can also be prescribed for medical use (e.g., to alleviate the pain related to diseases and surgeries). Thus, it is important to develop techniques to disambiguate whether a post retrieved is related to illicit drug use or not. In addition, hashtags are evolving. New hashtags are created constantly. This is especially true for illicit drug-related hashtags since drug users often want to evade detection. Thus, we also need to develop new techniques to keep tracking of the evolution of hashtags so that we will not miss new posts associated with new hashtags.

In this paper, we propose several techniques to identify and retrieve the social media posts related to drug abuse. We employ topic modeling to automatically disambiguate hashtags based on their topical context. We also propose a novel approach based on neural word embedding to track the evolution of hashtags.

II. Related Works

Addiction is a complex disorder with interacting factors, including personal traits, stress responsivity, social norms and attitudes, drug-induced neurobiological changes and co-morbidity. A growing number of studies have confirmed a strong association between physiological traits and addictive behaviors. Perry et al.[2] measure impulsivity with delayed discounting and find that impulsivity plays a role in several key transition phases of drug abuse. Terracciano et al. conduct a study involving 1102 participants and find a link between drug use and low *conscientiouness*, a personality trait related to a tendency to show self-discipline, act dutifully, and aim for achievement. [4] reveals risk factors leading to an addiction

[1] http://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

vulnerability such as age, sex/hormonal status, impulsivity, sweet-liking, novelty reactivity, proclivity for exercise, and environmental impoverishment. Additionally, some environmental and social factors are linked to drug addiction such as neighborhood environment[5], family environment [6], [7] and social norms[8], [9].

Traditionally, studies on drug use rely on surveys or interviews with a limited number of people. The advent of social media makes a large volume of user data available, which includes demographic information (age, gender, education etc.) and social activities (follower, following, like etc.). So far, social media data have been used in many applications such as personal trait analysis [10], community and event detection [11], influenza trend detection [12], crime monitoring [13], drug use pattern detection [14], [15]. To support the above research, frequently, the crucial first step is to identify and retrieve social media posts that are related to a particular topic or event.

Many studies show social media tags are accurate content descriptors[16][17]. In [18], the patterns extracted from social media tags are used to retrieve Flickr images of geographical landmarks. In [19], Chen et al. analyze temporal and locational distributions of tags to identify their corresponding events. In addition to tags, a wide variety of social relation types can be used to improve performance in many information retrieval tasks. For example, Agichtein et al. [20] exploit community feedback to build a classifier to identify high quality content. In [21], temporal, social and topical features are used to train a event classifier to distinguish event and non-event posts in Twitter. Yahia et al. [22] present network-aware search to incorporate social behavior into search. And a bipartite graph model is utilized to retrieve web videos[23]. So far, the data for social media-based studies are mainly collected using a keyword or hashtag-based search APIs. However, most of these methods do not take the ambiguity and the dynamics of the hashtags into consideration. Here, we focus on developing new technologies to address these issues.

III. Methodology

A. Data Collection

For this study, we focus on the content from Instagram. Similar approaches can be used to retrieve posts from other social media platforms such as Twitter and Facebook. To obtain an initial sample of drug-related posts in Instagram, we collected the official and street names of commonly abused drugs from the National Institute on Drug Abuse (NIDA) website [2]. These drug names are then used to retrieve posts from Instagram using its search API. Our current dataset includes 137,955 posts retrieved using 371 hashtags. After filtering out redundant posts, we have 116,885 posts in total. We also annotated a small number of posts for evaluation. Each post is assigned one of the four labels: 1.medical use of drugs 2. illicit use of drugs 3. not related to drugs, 4. not sure. After removing the posts in foreign languages or with only emojis, the annotated dataset includes 255 posts. Figure 1 shows the distribution of each type of annotations in this dataset. Among the 255 annotated posts, 171 ($> 60\%$) discusses not drug-related topics. In all drug related posts, only 4 posts ($< 2\%$) are related to medical use of drugs and 62 posts (24%) related to illicit use. Finally, 18($< 10\%$) of the posts are labeled as "Not Sure".

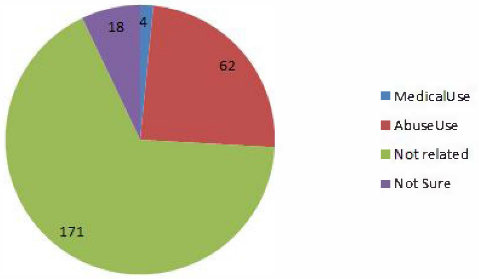


Fig. 1. Distributions of Annotated Posts

B. Identify Illicit Drug-related Social Media Posts

Latent Dirichlet Allocation (LDA) is a popular topic modeling technique. It can automatically discover the main topics in a text collection without any human supervision. LDA has been applied to various tasks such as automatic text summarization [24], [25], entity resolution [26], web spam classification [27], tag recommendation [28], topic extraction from the source code of software systems[29] and selectional preference modeling [30]. When applying LDA in topic analysis, it estimates the topic-term distribution $P(t|z)$ and the document-topic distribution $P(z|d)$ from an unlabeled corpus of documents given the two Dirichlet priors(α, β) and a fixed number (T) of topics[28].

We run our experiments using an online version of LDA[31], which is implemented in the Python Genism Library[3]. It allows new documents to update the model without running through the entire corpus. It can aid in processing large datasets in real time. In our experiments, we first remove stopping words from posts, then filter out small posts whose sizes are less than 3 words. We select three different T : 10, 20, 30. The two Dirichlet priors α, β default to a symmetric $1.0/T$ prior. Table I shows the top 10 topic keywords in three topics learned by LDA. As you can see, 1 and 13 are drug-related topics and 16 is not drug-related. Moreover, Topic 1 includes mostly drug names while Topic 13 has both drug names and general medical terms (e.g., patient, recovery, medical and clinic). Since there could be multiple topics in each drug abuse related post, we employ clustering to group the posts with similar topic distributions together. Based on the results, we identify one or more clusters that are related to the illicit use of drugs. In our experiment, each document is represented by the latent topic distribution inferred by LDA. Then we

[2] https://www.drugabuse.gov

[3] https://radimrehurek.com/gensim/install.html

TABLE I
SOME LATENT TOPICS BY TOP 10 TERMS

$T = 30$		
Topic 1	Topic 13	Topic 16
morphine	medical	lol
opiate	recovery	friend
addict	goodfellas	summer
fentanyl	dancefever	dj
xanax	sleep	laugh
percocet	patient	joke
oxycodone	clinic	love
lidocaine	stock	sun
painkill	opioid	ride
opium	problem	travel

use K-means clustering to group posts with similar topic distributions together. We have experimented with different Ks ($K = \{5, 6, 7, 8, 9, 10\}$). Each cluster is represented by a centroid which is the average across all the documents in the cluster. For model parameters, we empirically set $T = 30$ and $K = 5$ for a slightly better performance (Different values of T and K do not seem to change the results much). By inspecting the top 20 most representative posts in each cluster, we chose one of the five clusters as the illicit drug cluster. We evaluate the quality of our method with the small ground truth data set we mentioned earlier. After removing the posts annotated as ”Not Sure”, our system is able to predict drug-abuse related posts with 78.1% accuracy. Using the above method , here is a post which is identified as illicit drug use-related

1) Fentanyl overdoses have been declared a public health emergency in BC. Public policy should reflect the reality of drug use in our communities, not outdated Reaganesque ”War on Drugs” thinking. The system is failing, we have to do better. #fentanyl, #fentanylkills ...

In contrast, here are two posts that are not considered drug abuse-related although they are associated with drug-related hashtags such as #morphine and #ChinaTown.

2) The manual describes practical points to hypodermic injection including a list of hypodermic medications, administration guidelines, and their therapeutic effects. #physician #doctor #md #nurse #syringe #hypodermic-syringe #morphine #anesthesia #ouch ...
3) Somewhere in #Brooklyn Cousin -shoot @ #chinatown - DM me for collaboration friends :).

C. Detection of New Illicit Drug-related Hashtags

Due to the illegality of drug abuse, people who use, buy and sell illicit drugs use highly variable terminologies to avert law enforcement, concerned citizens, parents, and teachers. Some street names have entered the common vocabulary (e.g., identified by NIDA as a street name), while countless others remain unknown. Every day, the list of these terminologies increases and changes. Street names and terminologies used in activities related to drugs are constantly evolving. What is used today may become obsolete tomorrow. Constant changes in the vernacular helps drug users evade detection by others. This is why it is very important to detect these constantly changing terminologies in order to monitor drug abuse related posts in social media. In this paper we propose a novel approach to find terminologies related to drug abuse using semantic word embeddings.

Word embeddings map words to dense vectors of real numbers. In this representation, semantically similar words have similar vectors. Word embeddings are previously used to detect drug adverse effect [32], [33], sentiment classification [34], [35], name entity recognition [36], [37]. We have trained a word embedding model on Instagram posts and used these embeddings to derive new terminologies related drug abuse.

1) Neural Word Embedding

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing where words or phrases from the vocabulary are mapped to vectors of real numbers. A word embedding $W : words \rightarrow R_n$ is a parameterized function mapping words in some language to high-dimensional vectors (perhaps 200 to 500 dimensions). Typically, the function is a lookup table, parameterized by a matrix, θ, with a row for each word: $W_\theta(w_n) = \theta_n$. W is initialized to have random vectors for each word. Then the algorithm goes over a large text corpus and at every step, observes a target word and its context (neighboring words within a window). The target word's vector and the context words' vectors are then updated to bring them close together in the vector space (and therefore increase their similarity score). Other vectors (all of them, or a sample of them) are updated to become less close to the target word. After observing many such windows, the vectors become meaningful, yielding similar vectors to semantically related words.

Methods to generate this mapping include neural networks, dimensionality reduction on the word co-occurrence matrix, probabilistic models and explicit representation in terms of the context in which words appear. Among them, the most effective one being neural word embedding which uses neural network to generate the mapping. There are a few typical models for learning the vector representations of words using neural network.

Feedforward Neural Net Language Model (NNLM) : The probabilistic feedforward neural network language model has been proposed in [38]. It consists of input, projection, hidden and output layers. At the input layer, N previous words are encoded using 1-of-V coding, where V is the size of the vocabulary. The input layer is then projected to a projection layer P that has dimensionality NxD, using a shared projection matrix. As only N inputs are active at any given time, composition of the projection layer is a relatively cheap operation. The NNLM architecture is however complex for the computation between the projection and the hidden layer, as values in the projection layer are dense.

Recurrent Neural Net Language Model (RNNLM) : Recurrent neural network based language model has been proposed to overcome certain limitations of the feedforward NNLM in [39]. The RNNLM model has input, hidden and output layer. It does not have projection layer. In this model, recurrent matrix connects a hidden layer to itself using time-delayed connections. This allows the recurrent model to form some kind of short term memory, as information from the past can be represented by the hidden layer state that gets updated based on the current input and the state of the hidden layer in the previous time step.

Continuous Bag-of-Words Model : This model has been proposed in [40]. It tries to predict the current word based on its context. This model is similar to the feedforward NNLM, where the non-linear hidden layer is removed and the projection layer is shared for all the words (not just the projection matrix); thus, all the words get projected into the same position (their vectors are averaged). This model is called bag-of-words model as the order of words in the history does not influence the projection.

Skip-gram Model : This model has been also proposed in [40]. This model uses the current word to predict words in the target context. More precisely, each current word is used as an input to a log-linear classifier with a continuous projection layer. It is used to predict words within a certain range before and after the current word. Increasing the range improves quality of the resulting word vectors, but it also increases the computational complexity.

Among these four word embedding models, it has been shown in [40] that Skip-gram models works best in capturing semantic information from text.

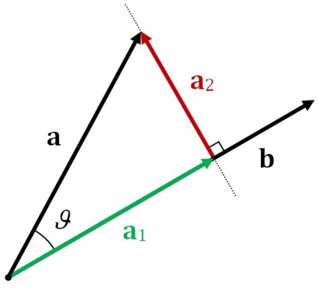
2) Method

We used word2vec skipgram model[40] to train word embeddings using the Instagram data set. Any word/hashtag that appeared less than 5 times in the data set was filtered out. Negative sampling was used while training. We have experimented with different feature vector size and context window size to test their influence on the quality of resultant vector. Feature vector size corresponds to the dimensionality of resultant word vectors and context wind size corresponds to the span of words in the text that is taken into account, backward and forward, when iterating through the words during model training. Optimization of these two parameters is very import to achieve good result. From our experiments we have found that combination of feature vector size 200 and context window size 5 works best in our data set. We have tried three different approaches to find hashtags and terminologies related to a given drug name based on our trained word embedding model.

Method 1: This method finds the most semantically related words of a drug name based on cosine similarity of word embeddings. Cosine similarity of two n-dimensional vectors A and B is defined as cosine angle between vector A and vector B. Word embedding of two semantically related words usually have large cosine similarity than any two random words. For example, the words ”man” and ”woman” will have larger cosine similarity than the words ”man” and ”tree”. Here we calculate cosine similarity of a drug name and all the words in our vocabulary. Then we sorted the words in ascending order of cosine similarity to find the most semantically related words of that drug.

Method 2: Vector rejection is useful to reject one particular meaning from a word. The rejection of **a** from **b** is the the vector component of **a** perpendicular to **b**. When a word has multiple meanings, vector rejection can be used to eliminate a meaning from a word. For example, the word ”bank” has two meanings: 1) ”river bank” and 2) ”the banking system”. We can extract a vector that means bank not in the sense of a depository institution by rejecting the vector of ”financial” from the vector of ”bank”. Figure 2 illustrates the concepts of projection and rejection in a vector space model.

Fig. 2. Projection of **a** on **b** is **a1**, and rejection of **a** from **b** is **a2**.



Many drugs especially prescription drugs have multiple usage (e.g., for medical use and for illicit use). As drug names appear in different contexts, when we try to find the most semantically related words of a drug, we found words from both contexts. With this method, we used vector rejection to reject medical usage-related meaning from a drug name and to keep drug abuse related meaning. After rejection, we use cosine similarity to find the most semantically related words to the drug name vector post rejection.

Method 3: Our third method is based on the hypothesis that two synonyms of a word have common semantically related words between them and two randomly chosen words will not have common semantically related words. We have used the number of common words as our criteria to determine whether two words/hashtags are related. To accomplish this, first we find the top 20 most semantically related words of a given drug name. Then we find the top 20 most semantically related words of each of these 20 words. Then we count how many common words shared between the top 20 semantically related words of the given drug name and the top 20 semantically related words of each of those top words. When we sort the words based on common word count, we get a new ranked list of the semantically related words/hashtags.

3) Experiment and Result

All the results are listed in Table III-C3.

Result of Method 1: We get some promising results using method 1. For example, among the top 20 most semantically related words of #marijuana, 6 of them (cannabis, herb, maryjane, pot, hash, ganja) are formally listed as the street names of marijuana in the website of National Institute of Drug Abuse (NIDA). The other terms are also mostly marijuana related. For example, although ”Kush” is a street name of marijuana, it is not listed in the NIDA website. #420 is the

TABLE II
TOP 20 DYNAMIC HASGTAGS GENERATED USING DIFFERENT METHODS

Rank	Method 1			Method 2	Method 3
	#marijuana	#fentanyl	#percocet	#fentanyl	#percocet
1	#cannabiscommunity	#hospital	#oxycodone	#oxycotin	#oxycodone
2	#cannabis	#midazolam	#xans	#lateralthinking	#hydros
3	#weed	#anesthesia	#roxicodone	#addrell	#oxys
4	#medicalmarijuana	#surgery	#oxys	#lortabs	#roxycodone
5	#herb	#medicine	#roxycodone	#hydrocodone	#kpins
6	#420	#ouch	#vicodin	#somas	#roxies
7	#maryjane	#wisdomtooth	#hydros	#nocros	#percocets
8	#pot	#katemine	#hydrocone	#alprazolam	#perces
9	#highlife	#painmeds	#percs	#vicodin	#oxycontin
10	#kush	#fentanylpatch	#roxy30s	#xanaxbars	#xans
11	#stonned	#demerol	#oxycontin	#ritalin	#hydrocodone
12	#cannabisconcentrates	#ativan	#dilaudid	#klonopin	#roxy30s
13	#weedporn	#painrelief	#opiates	#methadone	#roxicodone
14	#cannabisculture	#tramadol	#kpins	#xans	#rocicet
15	#trichomes	#sedation	#xanax	#Lol	#roxys
16	#weedstagram	#painpatch	#percocets	#benzos	#dilaudid
17	#ganja	#er	#roxies	#ativan	#vicodin
18	#drug	#trauma	#nacrotic	#qualitest	#xanax
19	#weedsociety	#icu	#roxycet	#actavis	#klonopin
20	#hash	#tocillectomy	#roxy	#pillpopper	#opiates

term that is used to represent marijuana consumption day/time. These words/hashtags can help us to find more posts related to marijuana use. For instance we searched Instagram using #cannabiscommunity and among top 50 posts we have found, 26 of them are marijuana related posts.

Result of Method 2: Among the top 20 words found for #fentanyl using method 1, some of them are general terms associated with medical usage. For example words like hospital, surgery, anesthesia, icu and medication. After we rejected the word ”#surgery” from ”#fentanyl” to reject the medical sense and obtain a new list of semantically related words. Now we can see that all medical usage related words are disappeared. So we can say that vector rejection can help us to reject the meaning of medical usage from drug names and find more relevant words for our purpose.

Result of Method 3: For this method first we found top 20 most semantically related words of #percocet. Then we found 20 most semantically related words of each of these words. Then we sorted them on the basis of counting common semantically related words. In this new list drugs those are not related to percocet (e.g. xans, dilaudid, vicotin, xanax, klonopin, opaites) rank lower than the original list.

D. Discussion

Our current dataset is rather small. With the current dataset, using semantic word embedding to extract drug abuse related terms from Instagram posts did not always provide expected result. This modest result can be explained by the small corpus size as well as the noisiness of the social media posts. The accuracy of word embedding models like word2vec largely depends on the number of time a word and its context words appear together. This is why the larger the dataset is, the better chance to have a particular drug name and a new street name appearing together. In general, this kind of word embedding models have billions of words. But our dataset currently has only 33225 words. Another problem while working with social media data is that the data is very noisy and we do not have much control over content quality. For instance, we collected all the posts based on the hashtags of a drug’s official and street names. We assumed that these hashtags will help us find drug related posts. But in reality, we found that even though a post contains hashtags of drug names or their street names, the actual posts retrieved are mostly none drug-related. So the data set is very sparse in terms of drug related posts. As in most posts, drug names, their street names and other drug abuse related terms didn’t appear in the same context, it is difficult to get street names or terms related to drug abuse using this method. We believe, with a big dataset, our methods will produce much better results. We will continue to collect more data to improve the performance.

IV. CONCLUSION

In this paper, we first propose a topic modeling-based approach for identifying drug abuse related posts. Our preliminary evaluation using a small ground truth dataset indicates that the system is able to retrieve drug related posts with 78.1% accuracy. Even with a small dataset, our word embedding model has shown promising results indicating that this method can be very useful to find new hashtags and terms related to drug abuse from social media. We can use the newly discovered hashtags to collect more drug abuse related posts from social media. As drug related terminologies are constantly evolving, our model allows us to retrieve and monitor drug abuse related posts in social media without being obsolete.

REFERENCES

[1] M. B. van den Bree, E. O. Johnson, M. C. Neale, and R. W. Pickens, “Genetic and environmental influences on drug use and

abuse/dependence in male and female twins," *Drug and alcohol dependence*, vol. 52, no. 3, pp. 231–241, 1998.

[2] J. L. Perry and M. E. Carroll, "The role of impulsive behavior in drug abuse," *Psychopharmacology*, vol. 200, no. 1, pp. 1–26, 2008.

[3] J. Szapocznik, G. Prado, A. K. Burlew, R. A. Williams, and D. A. Santisteban, "Drug abuse* in african american and hispanic adolescents: Culture, development, and behavior," *Annu. Rev. Clin. Psychol.*, vol. 3, pp. 77–105, 2007.

[4] M. E. Carroll, J. J. Anker, and J. L. Perry, "Modeling risk factors for nicotine and other drug abuse in the preclinical laboratory," *Drug and alcohol dependence*, vol. 104, pp. S70–S78, 2009.

[5] R. M. Crum, M. Lillie-Blanton, and J. C. Anthony, "Neighborhood environment and opportunity to use cocaine and other drugs in late childhood and early adolescence," *Drug and alcohol dependence*, vol. 43, no. 3, pp. 155–161, 1996.

[6] R. J. Cadoret, E. Troughton, T. W. O'Gorman, and E. Heywood, "An adoption study of genetic and environmental factors in drug abuse," *Archives of general psychiatry*, vol. 43, no. 12, pp. 1131–1136, 1986.

[7] D. A. Brent, "Risk factors for adolescent suicide and suicidal behavior: mental and substance abuse disorders, family environmental factors, and life stress," *Suicide and Life-Threatening Behavior*, vol. 25, no. s1, pp. 52–63, 1995.

[8] G. J. Botvin, "Preventing drug abuse in schools: Social and competence enhancement approaches targeting individual-level etiologic factors," *Addictive behaviors*, vol. 25, no. 6, pp. 887–897, 2000.

[9] E. Oetting and F. Beauvais, "Common elements in youth drug abuse: Peer clusters and other psychosocial factors," *Journal of Drug Issues*, vol. 17, no. 2, pp. 133–151, 1987.

[10] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI'11 extended abstracts on human factors in computing systems*. ACM, 2011, pp. 253–262.

[11] H. Sayyadi, M. Hurst, and A. Maykov, "Event detection and tracking in social streams." in *Icwsm*, 2009.

[12] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using twitter," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1568–1576.

[13] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 2012, pp. 1273–1276.

[14] A. Sarker, K. OConnor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter," *Drug safety*, vol. 39, no. 3, pp. 231–240, 2016.

[15] Y. Zhou, N. Sani, C.-K. Lee, and J. Luo, "Understanding illicit drug use behaviors by mining social media," *arXiv preprint arXiv:1604.07096*, 2016.

[16] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?" in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 2008, pp. 195–206.

[17] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 531–538.

[18] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: context and content in community-contributed media collections," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 631–640.

[19] L. Chen and A. Roy, "Event detection from flickr data through wavelet-based spatial analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 523–532.

[20] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008, pp. 183–194.

[21] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter." *ICWSM*, vol. 11, pp. 438–441, 2011.

[22] S. A. Yahia, M. Benedikt, L. V. Lakshmanan, and J. Stoyanovich, "Efficient network aware search in collaborative tagging sites," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 710–721, 2008.

[23] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang, "Web video topic discovery and tracking via bipartite graph reinforcement model," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 1009–1018.

[24] Y.-L. Chang and J.-T. Chien, "Latent dirichlet learning for document summarization," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 1689–1692.

[25] R. Arora and B. Ravindran, "Latent dirichlet allocation based multi-document summarization," in *Proceedings of the second workshop on Analytics for noisy unstructured text data*. ACM, 2008, pp. 91–97.

[26] I. Bhattacharya and L. Getoor, "A latent dirichlet model for unsupervised entity resolution." in *SDM*, vol. 5, no. 7. SIAM, 2006, p. 59.

[27] I. Bíró, J. Szabó, and A. A. Benczúr, "Latent dirichlet allocation in web spam filtering," in *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. ACM, 2008, pp. 29–32.

[28] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent dirichlet allocation for tag recommendation," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 61–68.

[29] G. Maskeri, S. Sarkar, and K. Heafield, "Mining business topics in source code using latent dirichlet allocation," in *Proceedings of the 1st India software engineering conference*. ACM, 2008, pp. 113–120.

[30] A. Ritter, O. Etzioni *et al.*, "A latent dirichlet allocation method for selectional preferences," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 424–434.

[31] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, 2010, pp. 856–864.

[32] A. Nikfarjam, A. Sarker, K. OConnor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*, p. ocu041, 2015.

[33] C. Wang, O. Singh, H. Dai, J. Jonnagaddala, T. R. Jue, U. Iqbal, E. Su, S. S. Abdul, and J. Li, "Nttmunsw system for adverse drug reactions extraction in twitter data," in *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, HI, USA*, 2016, pp. 4–8.

[34] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification." in *ACL (1)*, 2014, pp. 1555–1565.

[35] B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of sina weibo with word2vec," in *2014 IEEE International Congress on Big Data*. IEEE, 2014, pp. 358–363.

[36] A. Zirikly and M. Diab, "Named entity recognition for arabic social media," in *Proceedings of naacl-hlt*, 2015, pp. 176–185.

[37] N. Peng and M. Dredze, "Named entity recognition for chinese social media with jointly trained embeddings," in *Proceedings of EMNLP*, 2015, pp. 548–554.

[38] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[39] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.