Domain-specific Use Cases for Knowledge-enabled Social Media Analysis

Soon Jye Kho (✉), Swati Padhee (✉), Goonmeet Bajaj, Krishnaprasad Thirunarayan[0000−0002−7041−6963], and Amit Sheth[000−0002−0021−5293]

Kno.e.sis Center, Wright State University
Dayton 45435, Ohio, USA
{soonjye,swati,goonmeet,tkprasad,amit}@knoesis.org
http://knoesis.org

Abstract. Social media provides a virtual platform for users to share and discuss their daily life, activities, opinions, health, feelings, etc. Such personal accounts readily generate Big Data marked by velocity, volume, value, variety, and veracity challenges. This type of Big Data analytics already supports useful investigations ranging from research into data mining and developing public policy to actions targeting an individual in a variety of domains such as branding and marketing, crime and law enforcement, crisis monitoring and management, as well as public and personalized health management. However, using social media to solve domain-specific problem is challenging due to complexity of the domain, lack of context, colloquial nature of language and changing topic relevance in temporally dynamic domain. In this article, we discuss the need to go beyond data-driven machine learning and natural language processing, and incorporate deep domain knowledge as well as knowledge of how experts and decision makers explore and perform contextual interpretation. Four use cases are used to demonstrate the role of domain knowledge in addressing each challenge.

Keywords: Social Media, Machine Intelligence, Domain Knowledge, Knowledge Graph, Language Understanding, Mental Health Disorder, Emoji Sense Disambiguation, Drug Abuse Epidemiology, Implicit Entity Recognition

1 Introduction

As the saying goes, data is the new oil in the 21st century[1]. Data is described as an immense and valuable resource and extracting actionable insights for decision making is its value and has been the key to advancements in various domains. Many companies have realized this and have begun to treat their data as a valuable asset. This data is usually private and is not accessible to the public, resulting in a barrier for researchers. However, there is another source of continuously growing data that is publicly available and has been utilized by the research community for a variety of tasks – social media.

[1] https://goo.gl/YQq6pi

Social media serves as a virtual platform for users to share their opinions, report their daily life, activities, opinions, health, feelings, etc., and communicate with other users. In the year 2017, it has been reported that 81% of the public in America use some type of social media[2] and this contributes to the rapid generation of social media data. Twitter alone generates 500 million posts daily. Due to its volume and highly personalized nature, social media has been tapped by researchers for extracting useful individualized information. Various Natural Language Processing (NLP) techniques have been deployed for opinion mining [9, 27], sentiment analysis [10, 20], emotion analysis [17, 39], metadata extraction [1] and user group's characterization [28].

Just like crude oil is a raw material where its value can only be realized after the refinement process, social media data needs to be analyzed in order to gain actionable insights from it. The volume of social media data is a mixed blessing. From the view of data source, it is undoubtedly a bliss. The users share their views regarding a topic or an issue that can be very diverse and comprehensive in coverage, ranging from entertainment products to politics, and from personal health to religion. This explains why social media data has been used to gauge public perception on various issues and applications, such as brands uptake [36], election prediction[3] [5], disaster coordination [3, 30], public health issue such as monitoring depression [41], epidemiological and policy research related to epidemic [23], marijuana legalization [7] and drug abuse surveillance [4].

However, from the view of problem-solving, the content diversity and volume inherent in social media create significant practical challenges for extracting relevant information, as it is akin to searching for a needle in a haystack. To bridge the gap between the available raw data and the needed insights for decision making, researchers have exploited existing knowledge specific to the domain to fill in the gaps to relate pieces of data, determine their relevance to the problem at hand, and power an analytical framework. If the data was the crude oil which is to be refined to gain insights to solve a problem, then the domain knowledge is the instructions on how to use the refinery machines.

2 Challenges in Domain-specific Social Media Analysis

Social media analysis provides complementary insights for decision making. However, social media being an instance of Big Data creates several challenges for researchers to overcome before we can reap the promised fruits. Here, we discuss the major challenges in domain-specific social media analysis with respective examples in Table 1:

– Complex real-world domain - Real world problems are complex in that they involve multiple diverse factors. Understanding their mutual influences and interactions is non-trivial as exemplified many domains including healthcare, especially in the context of mental health disorder, its characteristics,

[2] https://goo.gl/i8n4Jm
[3] https://goo.gl/qYL2kv

causes, and progression[4]. Raw social media data often provides information at a basic level and seldom provides in an abstract actionable form that is necessary for solving the problems.
- Lack of context - Online users usually reveal their opinions and feelings [24] to their friends or in public assuming a shared understanding of the situation, which is implicit. Besides, social media such as Twitter impose word limitation further restraining the users from expressing more context. The presence of implicit context complicates a variety of NLP tasks such as entity disambiguation, sentiment analysis, and opinion mining.
- Colloquial nature of language - Due to the informal setting of social media, users tend to use language that is used in ordinary or familiar conversation. Instead of standard terms, users use slang terms and nicknames to refer an entity. Other than that, the language used on social media is prone to grammatical error, unconventional contractions, and spelling mistakes. This creates a challenge for entity recognition as social media data is collected based on defined keywords. If only standard terms are used to crawl for posts, it would fail to capture relevant information.
- Topic Relevance - Due to the dynamic or periodic nature of certain domains, the same context is relevant to different entities at different time periods. The domain knowledge, as well as their temporal aspect, is important to understand and disambiguate entities in such social media data.

3 Domain Knowledge for Social Media Analysis

Domain knowledge has been exploited by researchers to address the challenges mentioned above. Domain knowledge codifies an area of human endeavor or a specialized discipline [13]. It covers a broad range including facts, domain relationships, and skills acquired through experience or education. The use of domain knowledge in Web (semantic) applications such as search, browsing, personalization, and advertising was recognized and commercialized at the turn of the century [11, 32, 33] if not before, and has seen much wider usage when Google relied on its Knowledge Graph for its semantic search in 2013. Artificial Intelligence (AI) researchers have noted that "data alone is not enough" and that knowledge is very useful [8]. Use of knowledge to improve understanding and analysis a variety of textual and non-textual content, compared with the baseline machine learning and NLP techniques was discussed in [34].

As domain knowledge is normally acquired by humans through years of learning and experience, a domain expert is a scarce resource. Fortunately, there is an abundance of knowledge graphs (KG) [35] which are publicly available. The semantic web community has been putting major effort in producing large and cross-domain (e.g., Wikipedia and DBpedia) as well as domain-specific knowledge graphs (e.g., Gene Ontology, MusicBrainz, and UMLS) to codify well-circumscribed domain of discourse. Knowledge graphs [35] accessible on the web

[4] https://www.nimh.nih.gov/health/topics/index.shtml

Table 1. Examples of social media text that represent the challenges of social media analysis

Challenges	**Social Media Text**	**Problem**
Complex real-world domain	*I am disgusted with myself*	The text implies the user is having low self-esteem. Low self-esteem is one of the depressive symptoms, but it is untenable to diagnose the user with depression based on this individual text.
Lack of context	*WHOLE LOTTA* ⛽ A NG 💩 *GOIN ON* 😈 💯 🔫🔫	The gas pump emoji represents the letter 'G'. Without understanding the emojis and the hidden context, it is possible to infer the meaning of the text as: "WHOLE LOTTA GANG SHIT GOING ON".
Colloquial nature of language	*On sub ive actually felt nothing from +500mg doses for hours*	The term 'sub' is commonly used to refer Subutex, a brand name of buprenorphine. Using only standard drug name in data collection would fail to capture this tweet as relevant information for drug abuse.
Topic Relevance	*This new space movie is crazy. you must watch it!*	The term 'space movie' refers to different movies in different time periods. The domain knowledge and its temporal salient are crucial in accurately disambiguating the entities.

have been increasingly used to incorporate semantics in various applications, such as recommender system [15, 26], named entity disambiguation [37], document retrieval [31] and last but not least, social media analysis.

Knowledge graphs are large and creating domain knowledge that can support domain-specific social media analytics can be very challenging. To extract the domain-specific subgraphs from large knowledge graphs, Lalithsena [18] presented an approach by considering entity semantics [19] of the knowledge graphs. Existing approaches extract the subgraph by navigating up to a predefined number of hops in the knowledge graphs starting from known domain entities but this can lead to irrelevant entities and relationships in the knowledge graph. The approach proposed by Lalithsena [18] uses the entity semantics of the categories in the Wikipedia category hierarchy to extract domain-specific subgraphs (as shown in Figure 1). The approach combines the semantic type, lexical, and structural categories using probabilistic soft logic (PSL) [2] to incorporate different forms of evidence. The extracted domain-specific subgraph using the proposed approach reduced 74% of the categories from the simple n-hop expansion subgraph.

Researchers have been putting many efforts in advancing the extraction of domain-specific knowledge from large knowledge graphs. The extracted knowledge can be complementary to such domain-specific social media data analysis

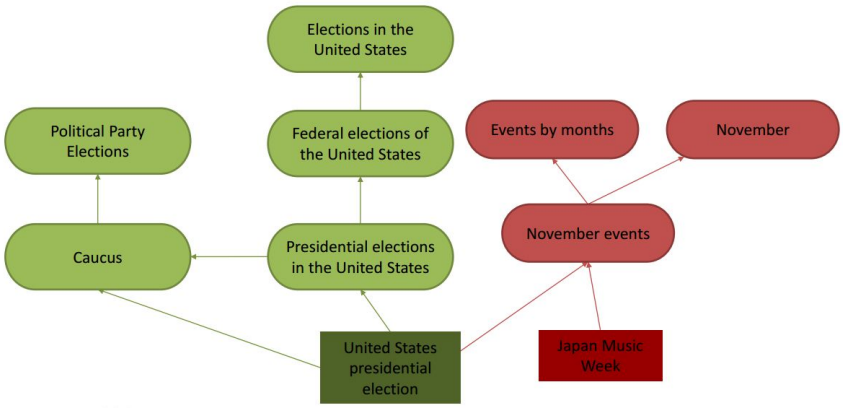

Fig. 1. Wikipedia category hierarchy for entity United States presidential election. Simple n-hop navigation would reach entity "Japan Music Week" which is irrelevant to the US presidential election. PSL based approach proposed by Lalithsena et al. successfully extracted a subgraph with more relevant entities (green color nodes).

(e.g., political election). In this article, we are going to demonstrate the role of domain knowledge in addressing the four major challenges and illustrate it using four appropriate use cases.

3.1 Addressing Domain Complexity - Use Case of Mental Health Disorder

Mental health disorders encompass a wide range of conditions that are characterized by a change of mood, thinking, and behavior[5]. These include but are not limited to obsessive-compulsive disorder, bipolar disorder, and clinical depression. Clinical depression is a common but debilitating mental illness that is prevalent worldwide. It affects more than 300 million people globally[6] and costs over $40 billion every year in the United States on depression treatment [6]. Diagnosis of depression in a patient requires a physician to consider a subset of predefined symptoms that last over a period of time. Typically, depression is detected in a primary care setting through patient health questionnaires[7] to screen for the presence and severity of depression.

Predefined symptoms of depression constitute the state-of-the-art domain knowledge curated by reputable medical doctors and are incorporated into tools like Patient Health Questionnaire (PHQ-9) [16]. PHQ-9 has been used by clinicians for screening, diagnosing, and measuring the severity of depression symptoms as defined by the American Psychiatric Association Diagnostic and Statis-

[5] https://www.nami.org/learn-more/mental-health-conditions
[6] http://www.who.int/mediacentre/factsheets/fs369/en/
[7] https://www.cdc.gov/mentalhealthsurveillance/

tical Manual (DSM-IV). The 9-item measurement tool contains a user-friendly response format, short administration time, and easy scoring [21].

However, studies have shown that these questionnaire-based methods are not accurate because of incomplete information and cognitive biases. First and foremost, there is no guarantee that patients can remember all accounts of depressive symptoms experienced over a certain period of time. Additionally, cognitive biases occur because of the way the questionnaire is phrased or administered, and this can prevent participants from giving truthful responses [12].

Contrary to our expectation that the social stigma associated with a clinically depressed patient will make them highly introverted and unlikely to share their condition publicly, there is a surprising amount of unsolicited sharing of depressive symptoms by users on social media. In fact, online social media provides a convenient and unobtrusive forum to voluntarily document their daily struggles with their mental health issues. This might be due to the anonymity that the social media provides, freeing them from the worries of people judging them. This type of passive monitoring can capture a user's condition with little burden on the user to report their emotions and feelings.

Social media text provides a source for researchers to monitor potential depressive symptoms, but it does not provide a good indication of whether a user is depressed or not. As mentioned above, depression is diagnosed based on the presence of 6 out of 9 symptoms over a period of two weeks as advocated by PHQ-9 questionnaire based diagnosis. Therefore, to model this type of domain practice through social media analysis, one should not just analyze the text in social media post disjointly. Instead, the analysis should perform jointly on every post of a user, as this portrays a more holistic picture of users' mental health condition.

Inspired by the scalability benefits and widespread use of PHQ-9 and social media analysis, Yazdavar et al. [41] emulates traditional observational cohort studies conducted through online questionnaires by extracting, categorizing, and unobtrusively monitoring different depressive symptoms. They developed a statistical model by modeling user-generated content in social media as a mixture of underlying topics evolving over time based on PHQ-9 symptoms.

To determine the textual cues that reflect the symptoms in each PHQ-9 criteria, Yazdavar et al. [41] generated a lexicon to capture each criterion and leveraged the lexicon as simple background knowledge. Furthermore, given the challenges of understanding the colloquial language on social media, Urban Dictionary[8] and the Big Huge Thesaurus[9] (a crowd-sourced online dictionary of slang words and phrases) were utilized. For expanding the lexicon, they used the synset of each of the nine PHQ-9 depression symptom categories. The consistency of the built lexicon has been vetted by domain experts. The final lexicon contains over 1,620 depression-related symptoms categorized into nine different clinical depression symptoms which are likely to appear in the tweets of individuals suffering from clinical depression (as shown in Table 2).

[8] https://www.urbandictionary.com/
[9] http://www.thesaurus.com/

Table 2. Few examples of lexicon terms and sample tweet for PHQ-9 symptoms.

PHQ-9	**Lexicon Terms**	**Sample Tweet**
Lack of Interest	"Couldn't care less" "don't want this" "Used to enjoy" "Zero motivation"	*I've not replied all day due to total lack of interest, depressed probs*
Feeling Down	"All torn up" "can't stop crying" "i'm beyond broken" "Shit day"	*i feel like i'm falling apart.*
Sleep Disorder	"Can't sleep" "sleep deprived" "sleeping pill" "crying to sleep"	*Night guys. Hope you sleep better than me.*
Lack of Energy	"Drained" "lassitude" "i am weak" "tired of everything"	*so tired, so drained, so done*
Eating Disorder	"Flat tummy" "hate my thighs" "love feeling hungry" "wanna be thinner"	*just wanna be skinny and beautiful*
Self Hate	"I am disgusting" "I'm a freak" "Waste of space" "Never good enough"	*I just let everyone down. why am I even here?*
Concentration Problems	"Overthinking" "short attention span" "can't pay attention" "brain dead"	*I couldn't concentrate to classes at all can't stop thinking*
Hyper/Lower Activity	"spazz" "paranoid" "unsettled" "i'm slow moving"	*so stressed out I cant do anything*
Suicidal Thoughts	"Ending my life" "overdosing" "my razor" "sleep forever"	*I want summer but then i don't... It'll be harder to hide my cuts.*

Using the lexicon, Yazdavar et al. identified users with self-reported symptoms of depression based on their profile descriptions. They then formulated a hybrid solution and conducted a temporal analysis of user-generated content on social media for capturing mental health issues. This use of domain knowledge proved valuable in creating a tool with an accuracy of 68% and a precision of 72% for capturing depression symptoms per user over a time interval.

To summarize, using the domain knowledge in the form of PHQ-9 categories and associated lexicons, they defined textual cues that could reflect depressive symptoms and use these cues in mining depression clues in social media data. They also uncovered common themes and triggers of depression at the community level. Not only emulating the PHQ-9 simplifies the domain complexity for understanding depression through social media, but it also provides a more intuitive interpretation that can be used to complement physicians in detecting depression.

3.2 Addressing Lack of Context - Use Case of EmojiNet

Users share their opinions, activities, locations, and feeling but they seldom express the motivation or intention behind their sharing. This is due to the shared understanding between friends, unwillingness of sharing too much detail, or word limitation by some platform (Twitter only allowed a maximum word count of 140 before they increased it to 280). Lack of context comprises one of the major challenges in social media analysis since context is critical in performing entity disambiguation, sentiment and emotion analysis.

Wijeratne et al. [40] exploit the use of emoji as a means to glean context to enhance understanding of a post. Emoji has been commonly used in social media and is extremely popular in electronic communication[10]. People use emoji to show whimsiness and describe emotions that are hard to articulate. As such, emoji provides an alternative and effective way to express intention, sentiment, and emotion in a post. A study showed that emoji polarity can improve the sentiment score and understand the meaning of emoji can improve the interpretation of a post [25].

However, a particular emoji can be ambiguous [40]. For example, 😂 (face with tears of joy) can be used for expressing happiness (using senses such as laugh and joy) as well as sadness (using senses such as cry and tears) [22]. In order to understand the sense conveyed through an emoji, Wijeratne et al. developed the first machine-readable sense inventory for emoji, EmojiNet[11] [40]. They integrated four openly accessible knowledge sources for emoji (i.e., Unicode Consortium, Emojipedia, iEmoji and The Emoji Dictionary) into a single dictionary of emoji.

EmojiNet represents a knowledge inventory of emoji where their senses and definitions are housed. This inventory is used for emoji sense disambiguation application which is designed to automatically learn message contexts where a

[10] https://goo.gl/ttxyP1
[11] http://emojinet.knoesis.org/

particular emoji sense could appear. Consider the above example of 😂 (face with tears of joy), the emoji sense disambiguation techniques could determine the sense of the emoji (happiness or sadness) based on the context in which it has been used. Using the EmojiNet, researchers are able to understand the meaning of emoji used in a message. Understanding the meaning of emoji hereby enriches the context that could potentially enhance applications that study, analyze, and summarize electronic communications.

3.3 Addressing Colloquial Nature of Language - Use Case of Drug Abuse Ontology

Opioids are a class of drugs that are chemically related and interact with opioid receptors in nerve cells. The medical opioid is available legally via prescription and the common medical use of an opioid is to relieve pain. Generally, it is safe to use them over a short-term and as prescribed by the doctor. However, due to its euphoric properties, opioid has been misused which can lead to overdose incidents [38].

The non-medical use of pharmaceutical opioids has been identified as one of the fastest growing forms of drug abuse in the United States. The White House Office of National Drug Control Policy (ONDCP)[12], in May 2011, launched America's Prescription Drug Abuse Crisis [14] initiative to curb prescription drug abuse problem, mainly through education and drug monitoring programs. This underscores the importance of determining new and emerging patterns or trends in prescription drug abuse in a timely manner.

Existing epidemiological data systems provide critically important information about drug abuse trends, but they are often time-lagged, suffering from large temporal gaps between data collection, analysis, and information dissemination. Social media offers a platform for users to share their experiences online and this information is useful in timely drug abuse surveillance.

To collect social media data that are related to drug use experience, it is important to recognize the mentions of drug-related entities in the tweets. However, entity recognition from social media data is difficult, due to grammatical errors, misspellings, usage of slang term and street names. For example, *Buprenorphine* might be referred to as *bupe*; marijuana concentrate products might be referred to as *dabs*, *earwax* or *hash oil* [7]. This colloquial nature of the language used in social media causes a decrease in the recall for capturing relevant information.

To address the challenge, Cameron et al. [4] manually curated Drug Abuse Ontology (DAO, *pronounced dao*) with the help of domain expert. They modeled the ontology using web forum posts, domain concepts, and relationships. DAO is the first ontology for prescription drug abuse and it is used for processing web forum posts along with a combination of lexical, pattern-based and semantics-based techniques. It contains the mapping of popular slang terms to the standard drug names and this provides a good base reference for drug entity recognition. This knowledge of slang term mapping and enrichment of vocabulary helps to

[12] https://www.whitehouse.gov/ondcp/

increase the recall for collecting relevant data, while the reduction in precision is prevented/remedied by using entity disambiguation.

3.4 Addressing Topic Relevance - Use Case of Movie Domain

Social media has been used to share information about entities such as movies and books. However, a novel aspect of this communication is that the entities are mentioned implicitly through their defining characteristics. Motivated by this, Perera et al. [29] have introduced and addressed this issue of identifying implicit references to an entity in tweets. An implicit entity is defined as an entity mentioned in a text where its name nor its synonym/alias/abbreviation is not present in the text [29]. For example, an entity (movie 'Gravity') can be explicitly mentioned as '*The movie Gravity was more expensive to make than the Mars Orbiter Mission.*' It can also be mentioned implicitly as '*This new space movie is crazy. you must watch it!.*' They reported that implicit mention of movie entity is found in 21% of the tweets. This indicates that keyword-based data collection would fail to capture almost one-quarter of relevant information and increasing the recall would be unattainable without the use of contextual domain knowledge.

Contextual domain knowledge is very important for understanding implicit mentions of entities, but it is complicated by the dynamic nature of some domain (e.g., media, movie, and news). These domains generate a lot of new facts that remain significant only for a short period of time because those new relationships came into prominence due to a related event in the news. Considering the movie domain, movies remain popular only for a limited time, being eclipsed by new movies. In fall of 2013, the mention of 'space movie' referred to the movie 'Gravity' whereas, in fall of 2015, it referred to the movie 'The Martian'. The temporal salience of this dynamically-changing domain knowledge that reflects the loose association between the movie and its referenced property – in this context the 'space movie' – is crucial for correctly disambiguating and identifying such implicit mentions.

Perera et al. proposed an approach which takes into account the contextual knowledge (common phrases in the tweets relevant to the implicit entity) as well as the factual knowledge (common terms, entities, and relationships relevant to the implicit entity). First, they acquired factual knowledge from DBpedia and extract only relevant knowledge based on their joint probability value with the given entity type. Then, they obtained contextual knowledge from contemporary tweets that explicitly mention the entity. These two types of knowledge are used to create an entity model network (EMN) to reflect the topical relationships among domain entities at a certain time. The EMN identifies relevant domain entities that are relevant in that time period and use this knowledge to identify implicit entity mentions.

The combination of factual knowledge and contextual knowledge which takes temporal salience into account enables the system to recognize implicit entity from tweets. Perera et al. evaluated their approach for two domains, viz., movies and books, and showed that the use of contextual knowledge contributed to

an improved recall of 14% and 19% respectively while the temporal salience improved the accuracy of entity disambiguation task by 15% and 18% respectively.

4 Discussion

Social media platforms have provided people a vehicle for free expressions of opinions, feelings, and thoughts. Researchers and computer scientists have exploited this rich content and the social interactions among its users to enable a deeper understanding of real-world issues and the people's perspectives on them. The progress in natural language processing techniques and its customization to social media communication provided the initial impetus for the analysis.

Semantic Web, or Web 3.0, has emerged as a complementary area of research for machine understanding of web data. This is being harnessed to capture and represent aspects related to human intelligence and cognition, which play a crucial role in human learning by enabling a deeper understanding of content and context at a higher abstraction level.

We discussed the importance and significance of domain knowledge in designing and exploiting social media analysis framework. Similar to the roles of knowledge in human intelligence, there are two major roles of domain knowledge in social media analysis: language understanding and information interpretation. Social media, as a platform for human expression and communication, inherently contains ambiguities, gaps, and implicit references similar to human natural language communication that requires domain knowledge for understanding and extracting the context. For example, Section 3.3 illustrates the use of different slang terms referring to the same drug entity that can be captured and represented through the use of ontology. This approach enables researchers to capture relevant information and fuse different types of information to obtain actionable insights in the medical context.

However, having a deep language understanding is not always sufficient for solving real-world problems that require domain knowledge for understanding the situation before generating actions. This is clearly illustrated in Section 3.1, where domain knowledge is needed to interpret the information extracted from the text for diagnosing depression. Domain knowledge provides an intuitive analysis framework by emulating how human experts practice in real life.

In this article, we have discussed four major challenges to domain-specific social media analysis and have highlighted the need to go beyond data-driven machine learning and natural language processing. Table 3 summarizes the specific challenges addressed by the application of domain knowledge, along with the lines of how experts and decision makers explore and perform contextual interpretation, to garner actionable information and insights from social media data.

Table 3. Application of domain knowledge in addressing major challenges in social media analysis with examples of domain-specific use case.

Challenges	**Use Case**	**Application and Improvement**
Complex real-world domain	Mental health disorder	Using PHQ-9 in analyzing tweets to simulate the diagnosis practice of physician in diagnosing depression.
Lack of context	EmojiNet	Understanding the meaning of emoji and use it to enrich the context.
Colloquial nature of language	Drug abuse epidemiology	Utilize the mapping of slang terms to standard drug references to improve recall and coverage of relevant tweets collection.
Topic relevance	Implicit entity linking	Extract related factual knowledge and contextual knowledge of a movie from DBpedia. The extracted knowledge and their temporal aspect are then taken into accounts for recognizing and disambiguating the movie/book entity in tweets.

Acknowledgments. We would like to thank Sarasi Lalithsena, Shweta Yadav, and Sanjaya Wijeratna for their patient and insightful reviews. We would also like to acknowledge partial support from the National Science Foundation (NSF) award: CNS-1513721: "Context-Aware Harassment Detection on Social Media", National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02: "Trending: Social Media Analysis to Monitor Cannabis and Synthetic Cannabinoid Use", National Institutes of Health (NIH) award: MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression", and Grant No. 2014-PS-PSN-00006 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the U.S. Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice, NSF, NIH or NIDA.

References

1. Abel, F., Hauff, C., Houben, G.J., Stronkman, R., Tao, K.: Twitcident: fighting fire with information from social web streams. In: Proceedings of the 21st International Conference on World Wide Web. pp. 305–308. ACM (2012)
2. Bach, S.H., Broecheler, M., Huang, B., Getoor, L.: Hinge-loss markov random fields and probabilistic soft logic. Journal of Machine Learning Research **18**(109), 1–67 (2017)

3. Bhatt, S.P., Purohit, H., Hampton, A., Shalin, V., Sheth, A., Flach, J.: Assisting coordination during crisis: a domain ontology based approach to infer resource needs from tweets. In: Proceedings of the 2014 ACM conference on Web science. pp. 297–298. ACM (2014)
4. Cameron, D., Smith, G.A., Daniulaityte, R., Sheth, A.P., Dave, D., Chen, L., Anand, G., Carlson, R., Watkins, K.Z., Falck, R.: Predose: a semantic web platform for drug abuse epidemiology using social media. Journal of biomedical informatics **46**(6), 985–997 (2013)
5. Chen, L., Wang, W., Sheth, A.P.: Are twitter users equal in predicting elections? a study of user groups in predicting 2012 us republican presidential primaries. In: International Conference on Social Informatics. pp. 379–392. Springer (2012)
6. Craft, L.L., Perna, F.M.: The benefits of exercise for the clinically depressed. Primary care companion to the Journal of clinical psychiatry **6**(3), 104 (2004)
7. Daniulaityte, R., Nahhas, R.W., Wijeratne, S., Carlson, R.G., Lamy, F.R., Martins, S.S., Boyer, E.W., Smith, G.A., Sheth, A.: "time for dabs": Analyzing twitter data on marijuana concentrates across the us. Drug & Alcohol Dependence **155**, 307–311 (2015)
8. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM **55**(10), 78–87 (2012)
9. Ebrahimi, M., Yazdavar, A.H., Salim, N., Eltyeb, S.: Recognition of side effects as implicit-opinion words in drug reviews. Online Information Review **40**(7), 1018–1032 (2016)
10. Ebrahimi, M., Yazdavar, A.H., Sheth, A.: Challenges of sentiment analysis for dynamic events. IEEE Intelligent Systems **32**(5), 70–75 (2017)
11. Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., Sheth, A.: Context and domain knowledge enhanced entity spotting in informal text. In: International Semantic Web Conference. pp. 260–276. Springer (2009)
12. Haselton, M.G., Nettle, D., Murray, D.R.: The evolution of cognitive bias. The handbook of evolutionary psychology (2005)
13. Hirschfeld, L.A., Gelman, S.A.: Mapping the mind: Domain specificity in cognition and culture. Cambridge University Press (1994)
14. House, W.: Epidemic: Responding to america's prescription drug abuse crisis. White House, Washington, DC (2011)
15. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on twitter using a hierarchical knowledge base. In: European Semantic Web Conference. pp. 99–113. Springer (2014)
16. Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9. Journal of general internal medicine **16**(9), 606–613 (2001)
17. Kušen, E., Cascavilla, G., Figl, K., Conti, M., Strembeck, M.: Identifying emotions in social media: comparison of word-emotion lexicons. In: Future Internet of Things and Cloud Workshops (FiCloudW), 2017 5th International Conference on. pp. 132–137. IEEE (2017)
18. Lalithsena, S.: Domain-specific knowledge extraction from the web of data. Department of Computer Science and Engineering, Wright State University (2018)
19. Lalithsena, S., Perera, S., Kapanipathi, P., Sheth, A.: Domain-specific hierarchical subgraph extraction: A recommendation use case. In: Big Data (Big Data), 2017 IEEE International Conference on. pp. 666–675. IEEE (2017)
20. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)

21. Milette, K., Hudson, M., Baron, M., Thombs, B.D., Group*, C.S.R.: Comparison of the phq-9 and ces-d depression scales in systemic sclerosis: internal consistency reliability, convergent validity and clinical correlates. Rheumatology **49**(4), 789–796 (2010)
22. Miller, H., Thebault-Spieker, J., Chang, S., Johnson, I., Terveen, L., Hecht, B.: Blissfully happy" or "ready to fight": Varying interpretations of emoji. Proceedings of ICWSM **2016** (2016)
23. Miller, M., Banerjee, T., Muppalla, R., Romine, W., Sheth, A.: What are people tweeting about zika? an exploratory study concerning its symptoms, treatment, transmission, and prevention. JMIR public health and surveillance **3**(2) (2017)
24. Mukherjee, S., Malu, A., AR, B., Bhattacharyya, P.: Twisent: a multistage system for analyzing sentiment in twitter. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2531–2534. ACM (2012)
25. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one **10**(12), e0144296 (2015)
26. Ostuni, V.C., Di Noia, T., Di Sciascio, E., Mirizzi, R.: Top-n recommendations from implicit feedback leveraging linked open data. In: Proceedings of the 7th ACM conference on Recommender systems. pp. 85–92. ACM (2013)
27. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. vol. 10 (2010)
28. Pennacchiotti, M., Popescu, A.M.: Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 430–438. ACM (2011)
29. Perera, S., Mendes, P.N., Alex, A., Sheth, A.P., Thirunarayan, K.: Implicit entity linking in tweets. In: International Semantic Web Conference. pp. 118–132. Springer (2016)
30. Purohit, H., Castillo, C., Diaz, F., Sheth, A., Meier, P.: Emergency-relief coordination on social media: Automatically matching resource requests and offers. First Monday **19**(1) (2013)
31. Schuhmacher, M., Ponzetto, S.P.: Knowledge-based graph document modeling. In: Proceedings of the 7th ACM international conference on Web search and data mining. pp. 543–552. ACM (2014)
32. Sheth, A., Avant, D., Bertram, C.: System and method for creating a semantic web and its applications in browsing, searching, profiling, personalization and advertising (Oct 30 2001), uS Patent 6,311,194
33. Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., Warke, Y.: Managing semantic content for the web. IEEE Internet Computing **6**(4), 80–87 (2002)
34. Sheth, A., Perera, S., Wijeratne, S., Thirunarayan, K.: Knowledge will propel machine understanding of content: Extrapolating from current examples. In: Proceedings of the International Conference on Web Intelligence. pp. 1–9. WI '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3106426.3109448, http://doi.acm.org/10.1145/3106426.3109448
35. Singhal, A.: Introducing the knowledge graph: things, not strings. Official google blog (2012)
36. Smith, A.N., Fischer, E., Yongjian, C.: How does brand-related user-generated content differ across youtube, facebook, and twitter? Journal of interactive marketing **26**(2), 102–113 (2012)

37. Usbeck, R., Ngomo, A.C.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: Agdistis-graph-based disambiguation of named entities using linked data. In: International Semantic Web Conference. pp. 457–471. Springer (2014)
38. Vowles, K.E., McEntee, M.L., Julnes, P.S., Frohe, T., Ney, J.P., van der Goes, D.N.: Rates of opioid misuse, abuse, and addiction in chronic pain: a systematic review and data synthesis. Pain **156**(4), 569–576 (2015)
39. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.P.: Harnessing twitter" big data" for automatic emotion identification. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). pp. 587–592. IEEE (2012)
40. Wijeratne, S., Balasuriya, L., Sheth, A., Doran, D.: Emojinet: Building a machine readable sense inventory for emoji. In: International Conference on Social Informatics. pp. 527–541. Springer (2016)
41. Yazdavar, A.H., Al-Olimat, H.S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., Sheth, A.: Semi-supervised approach to monitoring clinical depressive symptoms in social media. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 1191–1198. ACM (2017)