Knowledge discovery and hypothesis generation from online patient forums: A research proposal

Anne Dirkson
Leiden Institute of Advanced Computer Science, Leiden University
Niels Bohrweg 1, Leiden, the Netherlands
a.r.dirkson@liacs.leidenuniv.nl

Abstract

The unprompted patient experiences shared on patient forums contain a wealth of unexploited knowledge. Mining this knowledge and cross-linking it with biomedical literature, could expose novel insights, which could subsequently provide hypotheses for further clinical research. As of yet, automated methods for open knowledge discovery on patient forum text are lacking. Thus, in this research proposal, we outline future research into methods for mining, aggregating and cross-linking patient knowledge from online forums. Additionally, we aim to address how one could measure the credibility of this extracted knowledge.

1 Introduction

In the biomedical realm, open knowledge discovery from text has traditionally been limited to semi-structured data, such as electronic health records, and biomedical literature (Fleuren and Alkema, 2015). Patient forums (or *discussion groups*), however, contain a wealth of unexploited knowledge: the unprompted experiences of the patients themselves. Patients indicate that they rely heavily on the experiences of others (Smailhodzic et al., 2016), for instance for learning how to cope with their illness on a daily basis (Burda et al., 2016; Hartzler and Pratt, 2011).

In recent years, researchers have begun to acknowledge the value of such knowledge from experience, also called experiential knowledge. It is increasingly recognized as complementary to empirical knowledge (Carter et al., 2013; Knottnerus and Tugwell, 2012). Consequently, patient forum data has been used for a range of health-related applications from tracking public health trends (Sarker et al., 2016b) to detecting adverse drug responses (Sarker et al., 2015). In contrast to other potential sources of patient experiences such as electronic health records or focus groups, patient forums offer uncensored and unsolicited experiences. Moreover, it has been found that patients are more likely to share their experiences with their peers than with a physician (Davison et al., 2000).

Nonetheless, so far, the mining of experiential knowledge from patient forums has been limited to the extraction of adverse drug responses (ADRs) that patients experience when taking prescription drugs. Yet, patient forums contain an abundance of valuable information hidden in other experiences. For example, patients may report effective coping techniques for side effects of medication. Nevertheless, automated methods for *open* knowledge discovery from patient forum text, which could capture a wider range of experiences, have not yet been developed.

Therefore, we aim to develop such automated methods for mining anecdotal medical experiences from patient forums and aggregating them into a knowledge repository. This could then be cross-linked to a comparable repository of curated knowledge from biomedical literature and clinical trials. Such a comparison will expose any novel information present in the patient experiences, which could subsequently provide hypotheses for further clinical research, or valuable aggregate knowledge directly for the patients.

Although hypothesis generation in this manner could potentially advance research for all patient groups, we expect it to be the most promising for patients with rare diseases. Research into these diseases is scarce (Aymé et al., 2008): their rarity obstructs data collection and for-profit industry considers this research too costly. Aggregation of data from online forums could spur the coordinated, trans-geographic effort necessary to attain progress for these patients (Aymé et al., 2008).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 64–73
Florence, Italy, July 28 - August 2, 2019.

Problem statement Patient experiences are shared in abundance on patient forums. Experiential knowledge expressed in these experiences may be able to advance understanding of the disease and its treatment, but there is currently no method for automatically mining, aggregating, cross-linking and verifying this knowledge.

Research question To what extent can automated text analysis of patient forum posts aid knowledge discovery and yield reliable hypotheses for clinical research?

Contributions Our main contributions to the NLP field will be: (1) methods for extracting of aggregated knowledge from patient experiences on online fora, (2) a method for cross-linking curated knowledge and complementary patient knowledge, and (3) a method for assessing the credibility of claims derived from medical user-generated content. We will release all code and software related to this project. Data will be available upon request to protect the privacy of the patients.

2 Research Challenges

In order to answer this research question, five challenges must be addressed:

- *Data Quality* Knowledge extraction from social media text is complicated by colloquial language, typographical errors, and spelling mistakes (Park et al., 2015). The complex medical domain only aggravates this challenge (Gonzalez-Hernandez et al., 2017).
- *Named Entity Recognition (NER)* Previous work has been limited to extracting drug names and adverse drug responses (ADRs). Consequently, methods for extracting other types of relevant entities, such as those related to coping behaviour, still need to be developed. In general, layman's terms and creative language use hinder NER from user-generated text (Sarker et al., 2018).
- *Automatic Relation Annotation* Relation extraction from forum text has been explored only for ADR-drug relations. A more open extraction approach is currently lacking. The typically small size of patient forum data and the subsequent lack of redundancy is the main challenge for relation extraction. Other challenges include determining the presence, direction and polarity of relations and normalizing relationships in order to aggregate claims.
- *Cross-linking with Curated Knowledge* In order to extract novel knowledge, the extracted knowledge should be compared with curated sources. Thus, methods need to be developed to build comparable enough knowledge bases from both types of knowledge.
- *Credibility of Medical User-generated Content* In order to assess the trustworthiness of novel, health-related claims from user-generated online content, a method for measuring their relative credibility must be developed.

3 Prior work

In this section, we will highlight the prior work for each of these research challenges. Hereafter, in section 4, we will outline our proposed approach to tackling them in light of current research gaps.

3.1 Data quality

The current state-of-the-art lexical normalization pipeline for social media was developed by Sarker (2017). Their spelling correction method depends on a standard dictionary supplemented with domain-specific terms to *detect* mistakes, and on a language model of generic Twitter data to *correct* these mistakes. For domains that have many out-of-vocabulary terms compared to the available dictionaries and language models, such as medical social media, this is problematic and results in a low precision for correct domain-specific words.

Besides improving data quality through spelling normalization, it is essential to identify which forum posts contain patient experiences before knowledge can be extracted from these experiences. Previous research into systematically distinguishing experiences on patient forums is limited to a single study on Dutch forum data (Verberne et al., 2019). They identified narratives using only lower-cased words as features. Furthermore, specialized classifiers for differentiating factual statements about ADRs and personal experiences of ADRs on social media have also been developed (e.g. Nikfarjam et al. (2015)). However, these are too specialized to be suited for identifying patient experiences in general.

3.2 NER on health-related social media

Named entity recognition on patient forums is currently restricted to the detection of ADRs to prescription drugs (Sarker et al., 2015). Leaman et al. (2010) were the first to extract ADRs from patient forum data by matching tokens to a lexicon of side effects compiled from three medical databases and manually curated colloquial phrases. As lexicon-based approaches are hindered by descriptive and colloquial language use (O'Connor et al., 2014), later studies attempted to use association mining (Nikfarjam and Gonzalez, 2011). Although partially successful, concepts occurring in infrequent or more complex sentences remained a challenge.

Consequently, more recent studies have employed supervised machine learning, which can detect inexact matches. The current state-of-the-art systems use conditional random fields (CRF) with lexicon-based mapping (Nikfarjam et al., 2015; Metke-Jimenez and Karimi, 2015; Sarker et al., 2016a). Key to their success is their ability to incorporate textual information. Information-rich semantic features, such as polarity (Liu et al., 2016); and unsupervised word embeddings (Nikfarjam et al., 2015; Sarker et al., 2016a), were found to aid the supervised extraction of ADRs. As of yet, deep learning methods have not been explored for ADR extraction from patient forums.

For subsequent concept normalization of ADRs i.e. their mapping to concepts in a controlled vocabulary, supervised methods outperform lexicon-based and unsupervised approaches (Sarker et al., 2018). Currently, the state-of-the-art system is an ensemble of a Recurrent Neural Network and Multinomial Logistic Regression (Sarker et al., 2018). In contrast to previous research, we aim to extract a wider variety of entities, such as those related to coping, and thus we will also extend normalization approaches to a wider range of concepts.

3.3 Automated relation extraction on health-related social media

Research on relation extraction from patient forums has been explored to a limited extent in the context of ADR-drug relations. Whereas earlier studies simply used co-occurrence (Leaman et al., 2010), Liu and Chen (2013) opted for a two-step classifier system with a first classifier to determine *whether* entities have a relation and a second to define it. Another study used a Hidden Markov Model (Sampathkumar et al., 2014) to predict the presence of a causal relationship using a list of keywords e.g. 'effects from'. More recently, Chen et al. (2018) opted for a statistical approach: They used the Proportional Reporting Ratio, a statistical measure for signal detection, which compares the proportion of a given symptom mentioned with a certain drug to the proportion in combination with *all* drugs. In order to facilitate more *open* knowledge discovery on patient forums, we aim to investigate how other relations than ADR-drug relations can be extracted.

3.4 Cross-linking medical user-generated content with curated knowledge

Although the integration of data from different biomedical sources has become a booming topic in recent years (Sacchi and Holmes, 2016), only two studies have cross-linked user-generated content from health-related social media with structured databases. Benton et al. (2011) compared co-occurrence of side effects in breast cancer posts to drug package labels, whereas Yeleswarapu et al. (2014) combined user comments with structured databases and MEDLINE abstracts to calculate the strength of associations between drugs and their side effects. We aim to develop cross-linking methods with curated sources that go beyond ADR-drug relations in order to extract divergent novel knowledge from user-generated text.

3.5 Credibility of medical user-generated content

As the Web accumulates user-generated content, it becomes important to know if a specific piece of information is credible or not (Berti-Equille and Ba, 2016). For novel claims, the factual truth can often not be determined, and thus credibility is the highest attainable.

So far, the approaches to automatically assessing credibility of health-related information on social media has been limited to three studies (Viviani and Pasi, 2017a). Firstly, Vydiswaran et al. (2011) used textual features to compute trustworthiness based on community support. They evaluated their approach using simulated data with varying amounts of invalid claims, defined as disapproved or non-specific treatments, e.g. paracetamol. Secondly, Mukherjee et al. (2014) developed a semi-supervised probabilistic graph that uses an expert medical database of known side effects as a ground truth to assess the credibility of rare or

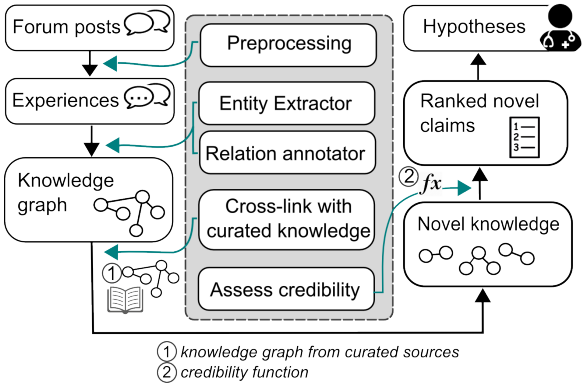

Figure 1: Proposed pipeline

unknown side effects on an online health community. Kinsora et al. (2017) was the first to not focus solely on accessing relations of treatments and side effects. They developed the first labeled data set of misinformative and non-misinformative comments from a health discussion forum, where misinformation is defined as 'medical relations that have not been verified'. By definition, however, the *novel* health-related claims arising from our knowledge discovery process will not be verified. Thus, so far, a methodology for assessing the credibility of novel health-related claims on social media is lacking. We aim to address this gap.

4 Proposed Pipeline

As can be seen in Figure 1, we propose a pipeline that will automatically output a list of medical claims from the knowledge contained in user-generated posts on a patient forum. They will be ranked in order of credibility to allow clinical researchers to focus on the most credible candidate hypotheses.

After preprocessing, we aim to extract relevant entities and their relations from only those posts that contain personal experiences. Therefore, we need a classifier for personal experiences as well as a robust preprocessing system. From the filtered posts, we will subsequently extract a wider range of entities than was done in previous research, such as those related to coping with adverse drug responses, medicine efficacy, comorbidity and lifestyle. Since patients with comorbities, i.e. co-occurring medical conditions, are often excluded from clinical trials (Unger et al., 2019), it is unknown whether medicine efficacy and adverse drug responses might differ for these patients. Moreover, certain lifestyle choices, such as diet, are known to influence both the working of medication (Bailey et al., 2013) and the severity of side effects. For instance, patients with the rare disease Gastro-Intestinal Stromal Tumor (GIST) provide anecdotal evidence that sweet potato can influence the severity of side effects.[1] These issues greatly impact the quality of life of patients and can be investigated with our approach. However, extending towards a more open information extraction approach instigates various questions. Could, for instance, dependency parsing be employed? Should a pre-specified list of relations be used and if so, which criteria should this list conform to? Which approaches and insights from other NLP domains could help us here?

Answering these questions is complicated by our consecutive aim to cross-link the patient knowledge with curated knowledge: the approach to knowledge extraction and aggregation needs to be similar enough to allow for filtering. A completely open approach may therefore not be possible. A key feature that impedes the generation of comparable data repositories is the difference in terminology. Extracting curated claims is also not trivial, as biomedical literature is at best semi-structured. Yet, comparable repositories are essential, as they will enable us to eliminate presently known facts from our findings.

Finally, we aim to automatically assess the credibility of these novel claims in order to output a ranked list of novel hypotheses to clinical researchers. Our working definition of credibility is the level of trustworthiness of the claim, or how valid the audience perceives the statement itself to be (Hovland et al., 1953). The development of a method for measuring credibility raises interesting points for discussion, such as: which linguistic features could be used to measure the credibility of a claim? And how could support of a statement, or lack thereof, by other forum posts be measured?

In the next two sections, we will elaborate, firstly, on initial results for improving data quality and, secondly, on implementation ideas for our NER and relation extraction system; and for our method for assessing credibility.

5 Initial results

To reduce errors in knowledge extraction, our research initially focused on improving data quality through (1) lexical normalization and (2) identify-

[1] https://liferaftgroup.org/managing-side-effects/

ing messages that contain personal experiences.[2]

Lexical normalization Since the state-of-the-art lexical normalization method (Sarker, 2017) functions poorly for social media in the health domain, we developed a data-driven spelling correction module that is dependent only on a *generic* dictionary and thus capable of dealing with small and niche data sets (Dirkson et al., 2018, 2019b). We developed this method on a rare cancer forum for GIST patients[3] consisting of 36,722 posts. As a second cancer-related forum, we used a subreddit on cancer of 274,532 posts [4].

For *detecting* mistakes, we implemented a decision process that determines whether a token is a mistake by, firstly, checking if it is present in a generic dictionary, and if not, checking for viable candidates. Viable candidates, which are derived from the data, need to have at least double the corpus frequency and a high enough similarity. This relative, as opposed to an absolute, frequency threshold enables the system to detect common spelling mistakes. The underlying assumption is that correct words will occur frequently enough to not have any viable correction candidates: they will thus be marked as correct. Our method attained an $F_{0.5}$ score of 0.888. Additionally, it manages to circumvent the absence of specialized dictionaries and domain- and genre-specific pretrained word embeddings. For *correcting* spelling mistakes, relative weighted edit distance was employed: the weights are derived from frequencies of online spelling errors (Norvig, 2009). Our method attained an accuracy of 62.3% compared to 20.8% for the state-of-the-art method (Sarker, 2017). By pre-selecting viable candidates, this accuracy was further increased by 1.8% point.

This spelling correction pipeline reduced out-of-vocabulary terms by 0.50% and 0.27% in the two cancer-related forums. More importantly, it mainly targeted, and thus corrected, medical concepts. Additionally, it increased classification accuracy on five out of six benchmark data sets of medical forum text (Dredze et al. (2016); Paul and Dredze (2009); Huang et al. (2017); and Task 1 and 4 of the ACL 2019 Social Media Mining 4 Health shared task[5]).

[2]Code and developed corpora can be found on https://github.com/AnneDirkson

[3]https://www.facebook.com/groups/gistsupport/

[4]www.reddit.com/r/cancer

[5]https://healthlanguageprocessing.org/smm4h/challenge/

Personal experience classification As research into systematically distinguishing patient experiences was limited to Dutch data with only one feature type (Verberne et al., 2019), we investigated how they could best be identified in English forum data (Dirkson et al., 2019a). Each post was classified as containing a personal experience or not. A personal experience did not need to be about the author but could also be about someone else.

We found that character 3-grams (F_1 = 0.815) significantly outperform psycho-linguistic features and document embeddings in this task. Moreover, we found that personal experiences were characterized by the use of past tense, health-related words and first-person pronouns, whereas non-narrative text was associated with the future tense, emotional support words and second-person pronouns. Topic analysis of the patient experiences in a cancer forum uncovered fourteen medical topics, ranging from surgery to side effects. In this project, developing a clear and effective annotation guideline was the major challenge. Although the inter-annotator agreement was substantial ($\kappa = 0.69$), an error analysis revealed that annotators still found it challenging to distinguish a medical fact from a medical experience.

6 Current and Future work

In the upcoming second year of the PhD project, we will focus on developing an NER and relation extraction (RE) system (Section 6.1). After that, we will address the challenge of credibility assessment (Section 6.2).

6.1 Extracting entities and their relations

For named entity recognition, we are currently experimenting with BiLSTMs combined with Conditional Random Fields. Our system builds on the state-of-the-art contextual flair embeddings (Akbik et al., 2018) trained on domain-specific data (Dirkson and Verberne, 2019). Our next step will be to combine these with Glove or Bert Embeddings (Devlin et al., 2018). We may also incorporate domain knowledge from structured databases in our embeddings, as this was shown to improve their quality (Zhang et al., 2019). The extracted entities will be mapped to a subset of preselected categories of the UMLS (Unified Medical Language System) (National Library of Medicine,

2009), as this was found to improve precision (Tu et al., 2016).

For relation extraction (RE), our starting point will also be state-of-the-art systems for various benchmark tasks. Particularly the system by Vashishth et al. (2018), RESIDE, is interesting as it focuses on utilizing open IE methods (Angeli et al., 2015) to leverage relevant information from a Knowledge Base (i.e. possible entity types and matching to relation aliases) to improve performance. We may be able to employ similar methods using the UMLS. Nonetheless, as patient forums are typically small in size, recent work in transfer learning for relation extraction (Alt et al., 2019) is also interesting, as such systems may be able to handle smaller data sets better. Recent work on few-shot relation extraction (Han et al., 2018) may also be relevant for this reason. Han et al. (2018) showed that meta-learners, models which try to *learn how to learn*, can aid rapid generalization to new concepts for few-shot RE. The best performing meta-learner for their new benchmark FewRel was the Prototypical Network by Snell et al. (2018): a few-shot classification model that tries to learn a prototypical representation for each class. We plan to investigate to what extent these various state-of-the-art systems can be employed, adapted and combined for RE in domain-specific patient forum data.

6.2 Assessing credibility

To assess credibility, we build upon extensive research into rumor verification on social media. Zubiaga et al. (2018) consider a rumor to be: "an item of circulating information whose veracity status is yet to be verified at time of posting". According to this definition, our unverified claims would qualify as rumors.

An important feature for verifying rumors is the aggregate stance of social media users towards the rumor (Enayet and El-Beltagy, 2017). This is based on the idea that social media users can collectively debunk inaccurate information (Procter et al., 2013), especially over a longer period of time (Zubiaga et al., 2016b). In employing a similar approach, we assume that collectively our users, namely patients and their close relatives, have sufficient expertise for judging a claim. Stances of posts are generally classified into supporting, denying, querying or commenting i.e. when a post is either unrelated to the rumor or to its veracity (Qazvinian et al., 2011; Procter et al., 2013). We plan to combine the state-of-the-art LSTM approach by Kochkina et al. (2017) with the two-step decomposition of stance classification suggested by Wang et al. (2017): comments are first distinguished from non-comments to then classify non-comments into supporting, denying, or querying. We will take into account the entire conversation, as opposed to focusing on isolated messages, since this has been shown to improve stance classification (Zubiaga et al., 2016a). We may employ transfer learning by using a pre-trained language model tuned on domain-specific data as input. Additional features will be derived from previous studies into rumor stance classification e.g. Aker et al. (2017).

For determining credibility, we plan to experiment with the model-driven approach by Viviani and Pasi (2017b), which was used to assess the credibility of Yelp reviews. They argue that a model-driven MCDM (Multiple-Criteria Decision Analysis) grounded in domain knowledge can lead to better or comparable results to machine learning if the amount of criteria is manageable on top of allowing for better interpretability. According to Zubiaga et al. (2018), interpretability is essential to make a credibility assessment more reliable for users. Alternatively, we may use interpretable machine learning methods, such as Logistic Regression or Support Vector Machines, similar to the state-of-the-art rumor verification system (Enayet and El-Beltagy, 2017). Besides stance, other linguistic and temporal features for determining credibility could be derived from rumor veracity studies e.g. Kwon et al. (2013); Castillo et al. (2011). We also plan to conduct a survey amongst patients in order to include factors they indicate to be important for judging credibility of information on their forum.

A challenge we foresee is the absence of a ground truth for the credibility of claims. To solve this, we could make use of the ground truth of claims that match curated knowledge through distant supervised learning and extrapolate our method to the unknown instances, comparable to the work by Mukherjee et al. (2014). Likewise, we could mirror Mukherjee et al. (2014) in our evaluation of the credibility scores: we could ask experts to evaluate ten random claims and the ten most credible as determined by our method.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria. INCOMA Ltd.

Christoph Alt, Marc Hbner, and Leonhard Hennig. 2019. Improving Relation Extraction by Pre-trained Language Representations. In *Automated Knowledge Base Construction 2019*, pages 1–18.

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 344–354.

Sgolne Aymé, Anna Kole, and Stephen Groft. 2008. Empowerment of patients: lessons from the rare diseases community. *The Lancet*, 371(9629):2048–2051.

David G Bailey, George Dresser, and J Malcolm O Arnold. 2013. Grapefruit-medication interactions: forbidden fruit or avoidable consequences? *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 185(4):309–16.

Adrian Benton, Lyle Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E Leonard, and John H Holmes. 2011. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation HHS Public Access. *J Biomed Inform*, 44(6):989–996.

Laure Berti-Equille and Mouhamadou Lamine Ba. 2016. Veracity of Big Data: Challenges of Cross-Modal Truth Discovery. *ACM Journal of Data and Information Quality*, 7(3):12.

Marika H.F. Burda, Marjan Van Den Akker, Frans Van Der Horst, Paul Lemmens, and J. Andr Knottnerus. 2016. Collecting and validating experiential expertise is doable but poses methodological challenges. *Journal of Clinical Epidemiology*, 72:10–15.

Pam Carter, Roger Beech, Domenica Coxon, Martin J. Thomas, and Clare Jinks. 2013. Mobilising the experiential knowledge of clinicians, patients and carers for applied health-care research. *Contemporary Social Science*, 8(3):307–320.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA. ACM.

Xiaoyi Chen, Carole Faviez, Stphane Schuck, Agns Lillo-Le-Louët, Nathalie Texier, Badisse Dahamna, Charles Huot, Pierre Foulquié, Suzanne Pereira, Vincent Leroux, Pierre Karapetiantz, Armelle Guenegou-Arnoux, Sandrine Katsahian, Cdric Bousquet, and Anita Burgun. 2018. Mining patients' narratives in social media for pharmacovigilance: Adverse effects and misuse of methylphenidate. *Frontiers in Pharmacology*, 9.

Kathryn P. Davison, James W. Pennebaker, and Sally S. Dickerson. 2000. Who talks? The social psychology of illness support groups. *American Psychologist*, 55(2):205–217.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*.

Anne Dirkson and Suzan Verberne. 2019. Transfer learning for health-related Twitter data. In *Proceedings of 2019 ACL workshop Social Media Mining 4 Health (SMM4H)*.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2019a. Narrative Detection in Online Patient Communities. In *Proceedings of the Text2StoryIR'19 Workshop at ECIR*, pages 21–28, Cologne, Germany. CEUR-WS.

Anne Dirkson, Suzan Verberne, Gerard van Oortmerssen, Hans van Gelderblom, and Wessel Kraaij. 2019b. Lexical Normalization of User-Generated Medical Forum Data. In *Proceedings of 2019 ACL workshop Social Media Mining 4 Health (SMM4H)*.

Anne Dirkson, Suzan Verberne, Gerard van Oortmerssen, and Wessel Kraaij. 2018. Lexical Normalization in User-Generated Data. In *Proceedings of the 17th Dutch-Belgian Information Retrieval Workshop*, pages 1–4.

Mark Dredze, David A Broniatowski, and Karen M Hilyard. 2016. Zika vaccine misconceptions: A social media analysis. *Vaccine*, 34(30):3441–2.

Omar Enayet and Samhaa R. El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474, Vancouver, Canada. Association for Computational Linguistics.

Wilco W M Fleuren and Wynand Alkema. 2015. Application of text mining in the biomedical domain. *Methods*, 74:97–106.

Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O 'Connor, and Guergana Savova. 2017. Capturing the Patient's Perspective: a Review of Advances

in Natural Language Processing of Health-Related Text. *Yearbook of medical informatics*, pages 214–217.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Andrea Hartzler and Wanda Pratt. 2011. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. *Journal of medical Internet research*, 13(3):e62.

C. I. Hovland, I. L. Janis, and H. H. Kelley. 1953. *Communication and persuasion; psychological studies of opinion change*. Yale University Press, New Haven, CT, US.

Xiaolei Huang, Michael C. Smith, Michael Paul, Dmytro Ryzhkov, Sandra Quinn, David Broniatowski, and Mark Dredze. 2017. Examining Patterns of Influenza Vaccination in Social Media. In *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*.

Alexander Kinsora, Kate Barron, Qiaozhu Mei, and V G Vinod Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In *IEEE International Conference on Healthcare Informatics*.

J. Andr Knottnerus and Peter Tugwell. 2012. The patients' perspective is key, also in research. *Journal of Clinical Epidemiology*, 65(6):581–583.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 117–125, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jing Liu, Songzheng Zhao, and Xiaodi Zhang. 2016. An ensemble method for extracting adverse drug events from social media. *Artificial Intelligence in Medicine*, 70:62–76.

Xiao Liu and Hsinchun Chen. 2013. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums. In *Smart Health. ICSH 2013. Lecture Notes in Computer Science*, pages 134–150. Springer, Berlin, Heidelberg.

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2015. Concept Extraction to Identify Adverse Drug Reactions in Medical Forums: A Comparison of Algorithms. *CoRR ArXiv*.

Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. 2014. People on Drugs: Credibility of User Statements in Health Communities. In *KDD'14*, pages 65–74.

National Library of Medicine. 2009. UMLS Reference Manual.

Azadeh Nikfarjam and Graciela H Gonzalez. 2011. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. *AMIA Annual Symposium proceedings*, 2011:1019–26.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association: JAMIA*, 22(3):671–81.

Peter Norvig. 2009. Natural Language Corpus Data. In Jeff Hammerbacher Toby Segaran, editor, *Beautiful Data: The Stories Behind Elegant Data Solutions*, pages 219–242. O'Reilly Media.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. In *AMIA Annual Symposium proceedings*, volume 2014, pages 924–33. American Medical Informatics Association.

Albert Park, Andrea L Hartzler, Jina Huh, David W Mcdonald, and Wanda Pratt. 2015. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. *J Med Internet Res*, 17(212).

Michael J Paul and Mark Dredze. 2009. A Model for Mining Public Health Topics from Twitter. Technical report, Johns Hopkins University.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Lucia Sacchi and John H Holmes. 2016. Progress in Biomedical Knowledge Discovery: A 25-year Retrospective. *Yearbook of medical informatics*, pages 117–29.

Hariprasad Sampathkumar, Xue-Wen Chen, and Bo Luo. 2014. Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. *BMC Medical Informatics and Decision Making*, 14.

Abeed Sarker. 2017. A customizable pipeline for social media text normalization. *Social Network Analysis and Mining*, 7(1):45.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2016a. Social Media Mining Shared Task Workshop. In *Pacific Symposium Biocomputing*, pages 581–592.

Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016b. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety*, 39(3):231–240.

Edin Smailhodzic, Wyanda Hooijsma, Albert Boonstra, and David J. Langley. 2016. Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Services Research*, 16(1):442.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2018. Prototypical networks for few-shot learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Hongkui Tu, Zongyang Ma, Aixin Sun, and Xiaodong Wang. 2016. When MetaMap Meets Social Media in Healthcare: Are the Word Labels Correct? In *Information Retrieval Technology. AIRS 2016. Lecture Notes in Computer Science.*, pages 356–362. Springer, Cham.

Joseph M. Unger, Dawn L. Hershman, Mark E. Fleury, and Riha Vaidya. 2019. Association of Patient Comorbid Conditions With Cancer Clinical Trial Participation. *JAMA Oncology*, 5(3):326.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Suzan Verberne, Anika Batenburg, Remco Sanders, Mies van Eenbergen, Enny Das, and Mattijs S Lambooij. 2019. Analyzing empowerment processes among cancer patients in an online community: A text mining approach. *JMIR Cancer*, 5(1):e9887.

Marco Viviani and Gabriella Pasi. 2017a. Credibility in social media: opinions, news, and health informationa survey. *WIREs Data Mining Knowl Discov*, 7.

Marco Viviani and Gabriella Pasi. 2017b. Quantifier Guided Aggregation for the Veracity Assessment of Online Reviews. *International Journal of Intelligent Systems*, 32(5):481–501.

V G Vinod Vydiswaran, Chengxiang Zhai, and Dan Roth. 2011. Gauging the Internet Doctor: Ranking Medical Claims based on Community Knowledge. In *KDD-DMH*.

Feixiang Wang, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 8: Rumour evaluation using effective features and supervised ensemble models. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 491–496, Vancouver, Canada. Association for Computational Linguistics.

SriJyothsna Yeleswarapu, Aditya Rao, Thomas Joseph, Vangala Govindakrishnan Saipradeep, and Rajgopal Srinivasan. 2014. A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC medical informatics and decision making*, 14(13).

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec,improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):52.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016a. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2438–2448.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016b. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.