Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

An annotated data set for identifying women reporting adverse pregnancy outcomes on Twitter

Ari Z. Klein*, Graciela Gonzalez-Hernandez

University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 18 August 2020
Accepted 25 August 2020
Available online 31 August 2020

Keywords:
Natural language processing
Machine learning
Social media
Data mining
Pregnancy
Epidemiology

ABSTRACT

Despite the prevalence in the United States of miscarriage [1], stillbirth [2], and infant mortality associated with preterm birth and low birthweight [3], their causes remain largely unknown [4–6]. To advance the use of social media data as a complementary resource for epidemiology of adverse pregnancy outcomes, we present a data set of 6487 tweets that mention miscarriage, stillbirth, preterm birth or premature labor, low birthweight, neonatal intensive care, or fetal/infant loss in general. These tweets are a subset of 22,912 tweets retrieved by applying hand-written regular expressions to a database containing more than 400 million public tweets posted by more than 100,000 women who have announced their pregnancy on Twitter [7]. Two professional annotators labeled the 6487 tweets in a binary fashion, distinguishing those potentially reporting that the user has personally experienced the outcome ("outcome" tweets) from those that merely mention the outcome ("non-outcome" tweets). Inter-annotator agreement was $\kappa = 0.90$ (Cohen's kappa). The tweets annotated as "outcome" include 1318 women reporting miscarriage, 94 stillbirth, 591 preterm birth or premature labor, 171 low birthweight, 453 neonatal intensive care, and 356 fetal/infant loss in general. These "outcome" tweets can be used to explore patient experiences and perceptions of adverse pregnancy outcomes, and can direct researchers to the users' broader timelines—tweets

DOI of original article: 10.1016/j.yjbinx.2020.100076

* Corresponding author.
E-mail address: ariklein@pennmedicine.upenn.edu (A.Z. Klein).

https://doi.org/10.1016/j.dib.2020.106249

posted by a user over time—for observational studies. Our past work demonstrates the analysis of timelines for selecting a study population [8] and conducting a case-control study [9] of users reporting that their child has a birth defect. For larger-scale studies, the full annotated corpus can be used to train supervised machine learning algorithms to automatically identify additional users reporting adverse pregnancy outcomes on Twitter. We used the annotated corpus to train feature-engineered and deep learning-based classifiers presented in "A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes" [10].

Specifications Table

Subject	Health informatics
Specific subject area	Social media mining for studying adverse pregnancy outcomes
Type of data	Text
How data were acquired	The raw data were acquired from a database of public tweets [7]. Binary labels were then manually provided by two annotators.
Data format	Raw, analyzed
Parameters for data collection	Tweets were collected if they mention miscarriage, stillbirth, preterm birth/premature labor, low birthweight, or neonatal intensive care.
Description of data collection	Handcrafted regular expressions retrieved 22,912 tweets that mention adverse pregnancy outcomes from a database containing public tweets posted by women who have announced their pregnancy on Twitter [7]. Two professional annotators labeled 8109 of the 22,912 tweets (one random tweet per user) in a binary fashion, distinguishing those potentially reporting that the user has personally experienced the outcome from those that merely mention the outcome. A random sample of 80% (6487) of the annotated tweets was selected to train supervised machine learning algorithms for automatic classification.
Data source location	Various
Data accessibility	With the article
Related research article	A.Z. Klein, H. Cai, D. Weissenbacher, L.D. Levine, G. Gonzalez-Hernandez, A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes, Journal of Biomedical Informatics: X, Available online 8 August 2020, 100076.

Value of the Data

- These tweets can be used to explore patient experiences and perceptions of adverse pregnancy outcomes, and can direct researchers to the users' broader timelines—tweets posted by a user over time—for observational studies.
- Sources of data for studying adverse pregnancy outcomes, especially early pregnancy loss, are limited, so Twitter data can benefit pregnant women, researchers, and clinicians as a resource that can help inform prenatal care by providing opportunities to gain insights from patients and assess potential risk factors.
- For larger-scale studies, the annotations can be used to train supervised machine learning algorithms to automatically identify additional users reporting adverse pregnancy outcomes on Twitter.

1. Data Description

This data set contains 6487 tweets (one tweet per user) that mention adverse pregnancy outcomes, labeled by two professional annotators in a binary fashion to distinguish tweets potentially reporting that the user has personally experienced the outcome ("outcome" tweets) from tweets that merely mention the outcome ("non-outcome" tweets). The tweets annotated as "outcome" include 1318 women reporting miscarriage, 94 stillbirth, 591 preterm birth or premature labor, 171 low birthweight, 453 neonatal intensive care, and 356 fetal/infant loss in general. The raw tweets can be downloaded using a Python script[1] and the "tweets_input.txt" file (Supplementary Material), which contains the user ID, tweet ID, adverse pregnancy outcome(s), and class ("0" = "non-outcome", "1"″ = "outcome") for each of the 6487 tweets. Only tweets that are still publicly available will be downloaded.

The 6487 tweets were retrieved from a database [7] using 11 handcrafted regular expressions—search patterns that define matching text strings (Supplementary Material). Table 1 presents samples of (slightly modified) tweets in the data set, and total distribution of "outcome" and "non-outcome" tweets for each of the 11 query patterns. For some outcomes, multiple query patterns were used, and some of the tweets in the data set match multiple query patterns. To facilitate the collection of additional "outcome" tweets for larger-scale studies, we have also provided keywords and their lexical variants [11] (Supplementary Material) that can be used to query Twitter directly through the API. The regular expressions, then, could be applied to the tweets returned from Twitter, followed by deploying a classifier trained on our annotated data set.

2. Experimental design, materials, and methods

We handcrafted 11 regular expressions to retrieve tweets that mention adverse pregnancy outcomes, from a database containing more than 400 million public tweets posted by more than 100,000 users who have announced their pregnancy on Twitter [7]. These query patterns were designed to account for the various ways adverse pregnancy outcomes may be linguistically expressed on social media—for example, reporting a miscarriage or stillbirth through the use of *rainbow baby* (Pattern 2) or hashtags such as *#babyloss, #pregnancyloss, #iam1in4*, or *#waveoflight* (Pattern 9), learned through an iterative process of manually reviewing tweets matched by other query patterns [8]. Similarly, preterm birth, for example, may be expressed by the user referring to her baby as a *preemie* (Pattern 4), or by reporting that her baby was born at less than 37 weeks of gestation (Pattern 5) or more than three weeks early (Pattern 7).

The query patterns were also designed to account for informal writing mechanics in social media. We automatically generated Twitter-based misspellings [11] of the keywords in the query patterns, and incorporated them into the regular expressions. Considering the character limit of tweets, the regular expressions also permitted abbreviations and optional spaces and coordinating conjunctions—for example, a low birthweight reported as *4lbs 12oz* (Pattern 8). Similarly, the only mention of an adverse pregnancy outcome may be embedded in a hashtag, such as *#endstillbirths* (Pattern 3), so the regular expressions did not require word boundary matches for words that would not result in excessive noise. Finally, our implementation of the regular expressions ignored letter casing in the tweets.

Initially, the query patterns focused on the recall of "outcome" tweets. Given the relatively low prevalence of adverse pregnancy outcomes in the general population, and our related work [8] suggesting that they may be under-reported on Twitter, high-precision query patterns would result in a sparse representation of "outcome" tweets; however, many of the preliminary regular expressions would have led to a high degree of class-imbalanced data [12] for training machine learning algorithms to automatically detect "outcome" tweets. Thus, to balance recall and pre-

[1] https://bitbucket.org/pennhlp/twitter_data_download/src/master/

Table 1
Samples and frequency of “outcome” (+) and “non-outcome” (-) tweets matching 11 query patterns. For ethical considerations, the tweets are slightly modified and usernames and URLs are redacted. The bold text indicates the string matched by the regular expression.

Query	*Outcome*	*Samples*	*Class*	*Frequency*
1	Miscarriage	*I was 3 months when I had my miscarriage* & I hate getting asked how my pregnancy is going & having to explain	+	1318
		A few months ago *I would have had an energy drink to fight this fatigue. But noo… caffeine causes miscarriages*.	–	2047
2	Fetal/infant loss	3 years ago today *I announced my pregnancy, which didn’t go as planned. Now blessed with this beautiful girl #rainbowbaby*	+	315
		My sister talking about her rainbow baby makes me sad but so happy! We’re so blessed.😟	–	94
3	Stillbirth	Stress can be vary dangerous! *I just gave birth to my 22 week old still born* baby girl, so now I’ve lost it all! 😭😭😟	+	83
		We’re talking about ways to reduce preventable newborn deaths, especially ways to #endstillbirths at [URL]	–	436
4	Preterm birth/labor	@[username] *My son was preemie* and it was hard to keep him awake while feeding. It gets better though!	+	248
		I continue to get bad cramps, I’m praying my baby isn’t born prematurely.	–	411
5	Preterm birth/labor	Really? *My son was born at 36 weeks* because I had preeclampsia! That’s crazy!	+	175
		I have known several babies that were born at 24 weeks and all of them but one is alive and growing up fine.	–	303
6	Neonatal intensive care	@[username] *my blues were intense, dh had to call my midwife friend because I wouldn’t stop crying but my baby was in SCBU*!	+	453
		People are so judgmental. *I don’t care if I give birth at a "baby factory" if it has a level III nicu*, which is important to me	–	138
7	Preterm birth/labor	My baby had her owns plans and decided to *arrive 5 weeks early*. Born March 1 ❤ I am so in love!	+	174
		I had a dream that my baby was *born 2 months early* and weighed 10 pounds exactly. I hope this isn’t some psychic stuff.😒	–	139
8	Low birthweight	On December 13th we *welcomed our sweet baby girl! Weighs 4lbs 12oz* and 18 and a half inches long	+	166
		last ultrasound before she’s *born* ❤ *she weighs 5 pounds & 1 ounce*	–	79
9	Fetal/infant loss	Thinking about *my baby today because today was your #duedate #babyloss*	+	41
		My baby only had a 1% chance to be born alive, thinking about the other families who didn’t get their **1%** ❤ *#infantloss*	–	14
10	Low birthweight	It turns out I was losing my amnio fluid but I didnt want to waste anyones time. My son was *born at* **1.3** kg	+	5
		My friends little boy was *born very early @* **821** g is 6mths old today. Still less than my son’s birth weight what a tough lil guy!	–	12
11	Stillbirth	@[username] *I support you, when in similar situation I also continued my pregnancy, my baby born sleeping* 1/23/2015	+	11
		I just read about a women who gave birth at 22 weeks (baby born sleeping) and I’m scared bc I’m 22 weeks right now!!	–	1

cision, the final regular expressions required a reference to the user (e.g., *I, our*), a child (e.g., *daughter, baby*), or birth (e.g., *born, welcome*) preceding the mention of an adverse pregnancy outcome, while allowing any number of characters to occur between.

The regular expressions retrieved 22,912 tweets (ignoring retweets) posted by 8109 users, from which we selected one random tweet per user for manual annotation. Annotation guidelines (Supplementary Material) were developed to help two professional annotators distinguish “outcome” and “non-outcome” tweets. We used 482 of the 8109 tweets to calibrate the annotations and guidelines; all of the remaining 7627 tweets were annotated independently by both annotators, with inter-annotator agreement of $\kappa = 0.90$ (Cohen’s kappa), considered “almost perfect agreement” [13]. The first author of this paper resolved the disagreements through independent annotation. Upon resolving the disagreements, 3653 (45%) tweets were annotated as “outcome,” and 4456 (55%) as “non-outcome.”

A random sample of 80% (6487) of the 8109 annotated tweets was selected as a training set for automatic classification. A deep neural network classifier achieved a benchmark F_1-score of 0.88[2] (precision = 0.87, recall = 0.89) for automatically detecting “outcome” tweets [10]. More specifically, the classifier achieved an F_1-score of at least 0.82 for automatically detecting “outcome” tweets reporting each of the adverse pregnancy outcomes, demonstrating the utility of this data set for deploying classifiers to automatically identify additional users for large-scale studies of specific outcomes. Our past work demonstrates the way that identifying users at the tweet level enables the analysis of the users’ broader timelines—tweets posted by a user over time—for selecting study populations [8] and conducting observational studies [9].

3. Ethics statement

The Institutional Review Board (IRB) of the University of Pennsylvania reviewed the studies for which this data was collected and deemed them exempt human subjects research under category (4) of paragraph (b) of the US Code of Federal Regulations Title 45 Section 46.101 for publicly available data sources (45 CFR §46.101(b)(4)). The Twitter data presented in this article is being distributed in accordance with the Twitter Developer Policy (https://developer.twitter.com/en/developer-terms/policy), accessed August 18, 2020.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) [grant number R01LM011176]. The content is solely the responsibility of the authors and does not necessarily represent the views of the NIH or NLM. The authors would like to acknowledge the annotators of the data set: Karen O’Connor, M.S. in Biomedical Informatics from Arizona State University, and Alexis Upshur, B.S. in Biology from Widener University, with ICD coding certification. They are full-time, professional annotators in the Department of Biostatistics, Epidemiology, and Informatics at the University of Pennsylvania.

[2] The test data will be available as part of a shared task (https://healthlanguageprocessing.org/smm4h21/).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.106249.

References

[1] L. Ammon Avalos, C. Galindo, D.K. Li, A systematic review to calculate background miscarriage rates using life table analysis, Birth Defects Res. A Clin. Mol. Teratol. 94 (6) (2012) 417–423.

[2] M.F. MacDorman, E.C.W. Gregory, Fetal and perinatal mortality: United States, 2013, Natl. Vital Stat. Rep. 64 (8) (2015) 1–24.

[3] J. Xu, S.L. Murphy, K. Kochanek, B. Bastian, E. Arias, Deaths: final data for 2016, Natl. Vital Stat. Rep. 67 (5) (2018) 1–76.

[4] L. Regan, R. Rai, Epidemiology and the medical causes of miscarriage, Baillieres Best Pract. Res. Clin. Obstet. Gynaecol. 14 (5) (2000) 839–854.

[5] K. Wou, M.P. Ouellet, M.F. Chen, R.N. Brown, Comparison of the aetiology of stillbirth over five decades in a single centre: a retrospective study, BMJ Open 4 (6) (2014) e004635.

[6] R. Menon, Spontaneous preterm birth, a clinical dilemma: etiologic, pathophysiologic and genetic heterogeneities and racial disparity, Acta Obstet. Gynecol. Scand. 87 (6) (2008) 590–600.

[7] A. Sarker, P. Chandrashekar, A. Magge, H. Cai, A. Klein, G. Gonzalez, Discovering cohorts of pregnant women from social media for safety surveillance and analysis, J. Med. Internet Res. 19 (10) (2017) e361.

[8] A.Z. Klein, A. Sarker, H. Cai, D. Weissenbacher, G. Gonzalez-Hernandez, Social media mining for birth defects research: a rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter, J. Biomed. Inform. 87 (2018) 68–78.

[9] S. Golder, S. Chiuve, D. Weissenbacher, A. Klein, K. O'Connor, M. Bland, M. Malin, M. Bhattacharya, L.J. Scarazinni, G. Gonzalez-Hernandez, Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy, Drug Saf. 42 (3) (2019) 389–400.

[10] A.Z. Klein, H. Cai, D. Weissenbacher, L.D. Levine, G. Gonzalez-Hernandez, A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes, Journal of Biomedical Informatics: X. Available online 8 August 2020, 100076.

[11] A. Sarker, G. Gonzalez-Hernandez, An unsupervised and customizable misspelling generator for mining noisy health-related text sources, J. Biomed. Inform. 88 (2018) 98–107.

[12] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.

[13] A.J. Viera, J.M. Garrett, Understanding interobserver agreement: the kappa statistic, Fam. Med. 37 (5) (2005) 360–363.