A lexicon-based approach to predicting pregnancy-related medication mentions by Twitter users

Stephen R. Piccolo
Department of Biology, Brigham Young University, Provo, Utah, United States of America

***Abstract*—We sought to develop an automated approach for classifying social media posts ("tweets") by 212 pregnant users. Part of the *BioCreative VII - Track 3 challenge*, we sought to differentiate between tweets that mentioned a medication use versus those that did not. Despite our initial intention to use more sophisticated methods, a manual review of tweets that mentioned medications suggested that a lexicon-based approach would perform well and have the advantages of being simple and fast. We derived a lexicon based on the training and validation sets as well as "Pharmacologic Substance" terms from the National Cancer Institute Thesaurus. We removed terms from our lexicon that were commonly used in non-medication tweets, even if the terms had been mentioned in some medication tweets. When applied to the challenge test set, this solution resulted in a higher Strict F1 Score than the other solutions, despite its simplicity. This solution might be most useful as a filter, enabling researchers to identify candidate tweets before applying more sophisticated machine-learning techniques.**

Keywords—Social Media; Medication Detection; Text Classification; Biomedical Ontology; Lexicon

I. Introduction

The purpose of this project was to identify social-media posts of 280 or fewer characters ("tweets" from the Twitter platform) that mention medication use(s) by individuals who were pregnant or recently had been pregnant at the time of their posts. Being able to identify such tweets in an automated, objective manner has potential to guide public-health efforts because researchers will better be able to identify individuals who are taking specific medications and monitor the prevalence of such use. One challenge that makes this process difficult is that tweets by pregnant women mentioning medication use are rare; thus there is a severe imbalance between positive and negative tweets. A second challenge is that social-media users often use colloquial language and refer to medications using a variety of terms, including generic medication names as well as commercial names. Furthermore, they may misspell medication names or use colloquial language to describe their use.

Track 3 - Automatic extraction of medication names in tweets of the BioCreative VII challenge provided an environment in which researchers could evaluate methods for differentiating between tweets from pregnant users who mentioned medication use and tweets from the same users that did not make such mentions. As an aid to help researchers get started in these efforts, the track organizers created a baseline implementation that used a lexicon for prefiltering and either convolutional neural networks or Bidirectional Encoder Representations from Transformers (BERT) to train binary classifiers(1–3). In evaluating the baseline labeler, they tried three resampling heuristics——under-sampling, fine-tuning, and filtering—and found that the BERT classifier in combination with multiple heuristics was most effective when tested on data from the Fifth Social Media Mining for Health Applications Workshop & Shared Task (SMMHH)(4). The lexicon used in the baseline implementation was based on the RxNorm vocabulary(5) and used an ensemble classification approach to account for spelling variants, semantics, and long-range dependencies(6). The baseline implementation extended this lexicon with additional terms, especially generic treatment terms. In addition, they manually removed terms that they deemed to be ambiguous.

The data for this challenge came from 212 Twitter timelines. Tweets mentioning a medication were annotated with spans (the start and end positions of the medication terms). Challenge organizers separated the tweets into training, validation, and test sets. So that participants would face a “real-world” scenario, no effort was made to mitigate class imbalance. Participants were asked to develop an automated system that would ingest raw tweets, predict which tweets mentioned medication(s), and assign a span within each tweet that had been predicted to mention a medication. This paper summarizes the methods and findings from our approach, which ranked first in the challenge among authorized submissions, according to Strict F1 Score.

II. Analysis on Training and Validation Sets

Knowing that the baseline implementation had used a machine-learning approach, we initially intended to do the same. But first, we created a lexicon. We queried the training set for medication mentions and used those terms as our initial lexicon. We then searched for these medications in the validation set, creating a simple regular expression for each medication. The regular expressions had word boundaries around each medication term; we also escaped any parentheses or square brackets and converted each medication term to lower case. Before searching for matching tweets in the validation set, we removed any hyperlinks or Twitter handles from the tweets. In preliminary testing, we found that relatively short medication terms (often substrings of longer terms) were often false positives, so we sorted the lexicon in descending order based on the number of characters per medication term. After identifying a medication in a given tweet, we excluded that tweet from consideration. However, our process did allow for multiple matches because in some cases multiple medications were mentioned in the same tweet.

This process resulted in 82 false positives and 38 false negatives on the validation set. 52 of the false positives were for the words “shot” and “shots.” Each of these terms had been used in a single tweet from the training set to describe a pregnancy medication. But in the remaining cases, these terms were not used to describe pregnancy medications. Of the remaining false positives, 11 were “pill” or “pills” and 9 were “prenatal.” Although “prenatal” is sometimes used to refer to prenatal vitamins (which were mentioned 8 times in the training set), it is used in other contexts, such as “prenatal yoga,” “prenatal massage,” and “prenatal classes.” The remaining false positives were “birth control” (n = 2), “flu shot” (n = 4), “insulin” (n = 1), “nasal spray” (n = 1), “prenatal vitamins” (n = 1), “steroids” (n = 1), and “vaccines” (n = 1). Of the false negatives, only “tylenol #3” was missed twice; all others were single misses, including relatively obscure terms—such as “Bio-oil,” “follistim,” and “rennie”—and more common medications—such as “aspirin,” “acetaminophen,” and “heparin.” Some false negatives were due to medications that are commonly known (e.g., “sleeping medication,” “sleeping pill,” “inhaler”). To our knowledge, none of the false negatives were due to misspellings. Based on these findings, we shifted our strategy in two ways: 1) expanding and refining our lexicon, and 2) ignoring terms that should be considered positive in rare cases but more frequently should be considered negative.

To expand our lexicon, we searched for a biomedical ontology that would be comprehensive enough to overcome many of the false negatives that we had encountered but that would not introduce too many false positives. Using BioPortal(7), we searched for terms that we had identified as false negatives. These following ontologies frequently appeared in the search results: RxNorm(5), Ontology of Consumer Health Vocabulary (OCHV)(8), and National Cancer Institute Thesaurus (NCIT)(9). For each of these ontologies, we considered preferred terms as well as synonyms. For RxNorm, we removed any term that was shorter than 5 characters, longer than 30 characters, had only numbers, started with a non-alphabetical character, or had a forward slash in it. To avoid conflicts with regular-expression syntax, we replaced any “+” character or text that was surrounded in square brackets with an empty string. After these steps, 34,601 unique terms remained. Matching these terms against the validation tweets resulted in 2020 false positives and 72 false negatives. The OCHV provides colloquial terms for some medications not used in other ontologies. It also includes terms that are unrelated to medications. After including synonyms and performing similar filtering steps, we identified 142,749 unique terms. When we searched against the validation set, we found 25,191 false positives and 62 false negatives. Finally, we evaluated the NCIT. This choice might seem unusual because NCIT is a cancer vocabulary. However, NCIT has an expansive list of medical terms that are used in diverse settings and includes many synonyms. In an attempt to prevent false positives, we limited NCIT to terms that fell under the “Pharmacologic Substance” heading (an option that was not available for OCHV). This filtering step excluded some terms such as “inhaler” that are medical devices, but it helped to focus our search on medications. We also excluded terms shorter than 5 characters or longer than 30 characters. This process resulted in 44,110 unique terms. Matching against the validation set resulted in 410 false positives and 75 false negatives. Of the three candidate vocabularies, we used NCIT in our lexicon because it resulted in the smallest number of false positives with only slightly more false negatives. After combining these terms with the medications identified in the training set, the number of false positives increased to 479. In addition to “shot,” “pill,” and other commonly used words that we had observed previously, NCIT included terms like “other” (n = 247), “oxygen” (n = 3), “medicine” (n = 16), “caffeine” (n = 9), and “alcohol” (n = 12). However, the number of false negatives decreased to 31. With the exception of “tylenol #3” (n = 2), each false negative was unique.

Next, we added known positives from SMM4H to our lexicon. This step reduced the number of false negatives from 31 to 25. But it increased the number of false positives from 479 to 532.

Having reduced false negatives considerably, we focused on reducing false positives. We excluded each term in our lexicon that had been mentioned in the training set in a non-medication tweet more times than it had been mentioned in medication tweets. Removing these terms reduced the number of false positives from 532 to 27. The number of false negatives remained 25. The terms “flu shot” and “birth control” were falsely identified as positive in multiple tweets; all other false positives occurred a single time (Table 1). In some cases, these terms were relevant to pregnancy (e.g., “prenatal vitamins,” “birth control,” “zofran pump”). In other cases, the terms had been positive in the training set but were not relevant to pregnancy in validation tweets. Except for “zofran,” which was twice a false negative, all false-negative terms occurred once (Table 2), although two variants of a given concept were false negatives in some cases (e.g., “whooping cough injection” and “whooping cough vaccine”). In other cases, a given tweet was associated with a false positive and a false negative due to a mismatching span. In cases where a tweet included two medication terms, we sometimes identified one term correctly but not the other.

TABLE I. FALSE-POSITIVE TERMS FROM THE VALIDATON SET. URLS, HANDLES, AND SPECIAL CHARACTERS HAVE BEEN REMOVED.

Tweet Text	Matched Span
I'm sitting here eating mustard and baking soda. My chest is still on fire	baking soda
You had my at biotin infusion...the pretty gold packaging was just a bonus	biotin
Great news! No more hospital visits until Monday. unfortunately, it's for my birth control implant. I'm tired of being poked.	birth control
Do with other birth control. So going to start talking to my doctor to see how soon after birth I can get it & I mean soon because	birth control
Going through my old books tonight & found this favorite... Rainbow Brite is a lovable little girl who,	Brite
It's totally fine that I just ate 7 Reece's Peanut Butter Cups for lunch because I'm providing both protein & calcium to my fetus. Duh.	calcium

Only thing shocking is the no rape part: CeeLo Green accused of giving woman ecstasy but DA declines rape	ecstasy
Got my flu shot in my left arm today and I'm supposed to sleep on my left side.....rookie move. #fluless	flu shot
My arm hurts from this damn flu shot smh	flu shot
Flu shot. Flu shot.	Flu shot
This flu shot has my arm feeling like someone took a bat to it	flu shot
Hydrocodone bitartrate & Celebrex I'm so set	Hydrocodone
Got myself a new bff! I'm now sat on my bed on an insulin drip. Would've liked a little wander,	insulin
lanolin nipple cream is good if u want some	lanolin
She is currently on a morphine drip & hospice is taking over.Trying to get her into an end of life facility for as much comfort as possible.	morphine drip
Between my steroid nasal spray & the oversized monster I just drank I'm feeling a little tweeky.	nasal spray
The nightly pill arsenal... Acetaminophen, 3 different components to my prenatal vitamins, and	prenatal vitamins
My face has puffed out so much I literally look like a puffer fish	puffer
Today has been a great Friday at work far. brought me a Strawberry/Rhubarb Pie and I got ice cream too!!!! #happybelly	Rhubarb
My organic red raspberry leaf tea gets delivered today (AKA steroids for my lady parts) time to get this uterus ready to expell human life	steroids
I'm going to tell my doctor I want the pill or the patch, I'm just hoping I do well on them	the pill
Got my whooping cough vaccine yesterday and now my arm is sore.	vaccine
Different veggies replace vaccines, apparently. rt Kristin implores us to eat fermented foods.	vacccines
Last night I put Vicks Vapo rub on my lips instead of Blistex....tonight I double checked before slathering it on.	Vicks
I look like Vitamin C from the 90s lololol	Vitamin C
I worried my vitamin D was low but just discovered in today's that I can conquer it by eating 12 packs of Cheddar a day. Score!	vitamin D
I have a possible infection from my Zofran pump.. seriously. Not catching any breaks during this pregnancy apparently.	Zofran pump

TABLE II. FALSE-NEGATIVE TERMS FROM THE VALIDATON SET. URLS, HANDLES, AND SPECIAL CHARACTERS HAVE BEEN REMOVED.

Tweet Text	Correct Span
Bio-oil is good but poundland sell something very similar which I find works just as well. The packaging look v. similar.	Acetaminophen
These new anti-anxiety meds make me so sleepy Not sure if It will work out for me.. But I'm so tired of always switching meds	anti-anxiety meds
Holy #### I need my freaking anxiety meds	anxiety meds
Has anyone tried making a crushed Asprin + water paste for pimples? Curious if it works! #beautychat	Asprin
Bio-oil is good but poundland sell something very similar which I find works just as well. The packaging look v. similar.	Bio-oil
No first thing in the morning and still have this migraine!!! It's been now 4 days and nothing is working! Including a blood patch	blood patch
So I'm just about to drink castor oil again because at this point I'm tired	castor oil
Hydrocodone bitartrate & Celebrex I'm so set	Celebrex
Can't see bc put earache drops instead of eye drops into my eyes.	earache drops
IVF Update - First Injection of Menopur and Follistim:	Follistim
I swear every few months I am on an inhaler & steroids because my lungs never developed properly. It is NOT fun..	inhaler
Got myself a new bff! I'm now sat on my bed on an insulin drip. Would've liked a little wander,	insulin drip
She is currently on a morphine drip & hospice is taking over.Trying to get her into an end of life facility for as much comfort as possible.	morphine
I'm in charge of narcotics & domestic violence what the ######	narcotics
I am. We are going through 14pts a week & I'm pretty much the only person who has any. And orange Rennies. It's just blurgh.	Rennies
I thought you were sleeping.. I can't its not gonna happen I took sleeping medicine and everything.	sleeping medicine
Between my steroid nasal spray & the oversized monster I just drank I'm feeling a little tweeky.	steroid nasal spray
Got my Tdap at 37 weeks... Hope it wasn't too late. Article states 27-36 weeks.	Tdap
I get those contractions ALL. THE. TIME. My doctor Put me On Vastaril to Help settle them down. It helps a bit.	Vastaril
Last night I put Vicks Vapo rub on my lips instead of Blistex....tonight I double checked before slathering it on.	Vicks Vapo rub
Anyone have muscle relaxers or Vicodeine !??	Vicodeine
The whooping cough injection site has killed my arm.	whooping cough injection
Got my whooping cough vaccine yesterday and now my arm is sore.	whooping cough vaccine
I have a possible infection from my Zofran pump.. seriously. Not catching any breaks during this pregnancy apparently.	Zofran
Ugh. My next syringe change is going to be around 4 am. Whyyy. #LifeWithAZofranPump	Zofran

III. TEST-SET PREDICTIONS

Before making test-set predictions, we created a final lexicon based on a) medications identified in training-set tweets, b) medications identified in the SMM4H tweets, c) NCIT medications that had been mentioned in the validation set, and d) the term "inhaler," which we identified as a likely false negative via manual observation of the validation set. We excluded any of these terms that occurred in negative tweets from the training and validation sets more frequently than they occurred in positive tweets. Finally, we excluded additional medication terms that we had manually identified as likely false positives based on the validation set: "shot," "shots," "flu shot," "pill," "pills," "nasal spray," "prenatal," "muscle relaxer," "hydrocodone," "vitamin e," "vitamin," "zofran pump," "juice plus," "morphine drip." The resulting lexicon had 1355 terms and can be found at https://osf.io/ma672/.

Our approach resulted in an Overlapping F1 Score of 0.755 and a Strict F1 score of 0.705. These scores were slightly higher than the mean among competitors but slightly lower than the median. The Strict F1 score was highest among the competitors. Although detailed results are not available at the time of this writing, it appears that our lexicon-based approach was most successful at making exact matches but struggled with fuzzy matching.

After the competition ended, we made a revised submission, considering RxNorm terms as well. After filtering based on negative tweets, only "injection" was added to our lexicon. This submission resulted in an Overlapping F1 Score of 0.741 and a Strict F1 Score of 0.691. Both scores were slightly lower than we had attained without this additional lexicon term. The full code from our analysis can be found at https://osf.io/nxb7p/.

IV. DISCUSSION

Our contribution to this challenge uses a lexicon-matching approach. Despite its simplicity, this technique performed similarly or marginally better than alternative approaches on the test set. The challenge organizers stated that "this competition will be an opportunity to go beyond the lexical match approach" (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/). However, ancillary goals of natural language processing are simplicity and speed. In particular, speed is critical for a real-world implementation because of the huge volume of tweets created each day. Thus, as our solution progressed, we decided to focus on a lexical-matching approach, informed by an external vocabulary and both positive and negative examples. Our approach uses basic programming logic and does not require any machine-learning libraries. Thus, it would be relatively easy for others to understand and adopt. Due to the size of NCIT, we filtered terms from this vocabulary by identifying whether a given term was present in the validation set. This reduced the size of our lexicon to approximately 3% of its original size and enabled us to make our final predictions in a few minutes (using the R statistical package). This speed could be improved further using an alternative programming language and through additional optimization strategies.

As with a machine-learning approach, our technique requires many positive and negative examples. Our lexicon could be refined further through manual curation of pregnancy-related tweets, including the test set. Using external vocabularies is helpful, but a key challenge is to exclude terms that are most likely to cause false positives because these vocabularies are so comprehensive. In addition to using negative tweet examples, a helpful strategy might be to compare against corpora from non-tweet sources that are likely not to include medication mentions.

Although our approach performed reasonably well in this competition, we do not see it as a candidate for real-world implementation on its own. We see it as a solution that could be used in conjunction with other approaches. As with the baseline implementation, it might make sense to use a lexical-matching approach as a prefiltering step, perhaps using multiple external vocabularies to minimize false negatives. Having reduced the number of candidate tweets (and reducing class imbalance), a machine-learning approach that uses word embeddings might help to classify the remaining tweets more accurately than a lexical-matching approach.

ACKNOWLEDGMENT

We thank the organizers of the BioCreative VII Challenge, especially those who coordinated efforts for Task 3.

REFERENCES

1. Weissenbacher D, Rawal S, Magge A, Gonzalez-Hernandez G. Addressing Extreme Imbalance for Detecting Medications Mentioned in Twitter User Timelines. In: Tucker A, Henriques Abreu P, Cardoso J, Pereira Rodrigues P, Riaño D, editors. Artificial Intelligence in Medicine. Cham: Springer International Publishing; 2021. p. 93–102. (Lecture Notes in Computer Science).

2. Chellapilla K, Puri S, Simard P. High Performance Convolutional Neural Networks for Document Processing. In: Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft; 2006.

3. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs [Internet]. 2019 May; Available from: https://arxiv.org/abs/1810.04805

4. Klein A, Alimova I, Flores I, Magge A, Miftahutdinov Z, Minard A-L, et al. Overview of the fifth social media mining for health applications (# Smm4h) shared tasks at coling 2020. In: Proceedings of the fifth social media mining for health applications workshop & shared task. 2020. p. 27–36.

5. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011;18(4):441–8.

6. Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. J Am Med Inform Assoc. 2019;26(12):1618–26.

7. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: Ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 Jul;37(suppl_2):W170–3.

8. Amith M, Cui L, Roberts K, Xu H, Tao C. Ontology of Consumer Health Vocabulary: Providing a formal and interoperable semantic resource for linking lay language and medical terminology. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2019. p. 1177–8.

9. Sioutos N, Coronado S de, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics. 2007 Feb;40(1):30–43.