Contents lists available at ScienceDirect

Data in Brief

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

Data Article

A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities

Abeed Sarker*, Graciela Gonzalez

Division of Informatics, Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 29 February 2016
Received in revised form
5 November 2016
Accepted 15 November 2016
Available online 23 November 2016

ABSTRACT

In this data article, we present to the data science, natural language processing and public heath communities an unlabeled corpus and a set of language models. We collected the data from Twitter using drug names as keywords, including their common misspelled forms. Using this data, which is rich in drug-related chatter, we developed language models to aid the development of data mining tools and methods in this domain. We generated several models that capture (i) distributed word representations and (ii) probabilities of n-gram sequences. The data set we are releasing consists of 267,215 Twitter posts made during the four-month period—November, 2014 to February, 2015. The posts mention over 250 drug-related keywords. The language models encapsulate semantic and sequential properties of the texts.

Specifications Table

Subject area	*Biomedical informatics, data mining, natural language processing*
More specific subject area	*Social media mining for public health*
Type of data	*Text, Binary*

* Corresponding author.
E-mail address: abeed@upenn.edu (A. Sarker).

http://dx.doi.org/10.1016/j.dib.2016.11.056

How data was acquired	*Drug-related chatter was directly collected from Twitter using the Twitter Streaming API. Posts were retrieved using drug names as keywords. To address the issue of common misspellings in social media, we employed a phonetic spelling variant generator that automatically generates common misspellings for the drug names.*
Data format	*Raw, processed*
Experimental factors	*Twitter posts were collected in raw format. Only basic preprocessing such as lowercasing was performed prior to the generation of the language models.*
Experimental features	*The language models were generated from the raw Twitter data after basic preprocessing. A neural network based technique is used to learn the distributed word representation models. The sequential language model is learned by computing probabilities of word n-gram sequences. The utilities of the two sets of models were verified via preliminary experiments of adverse drug reaction detection and text classification, respectively.*
Data source location	*N/A*
Data accessibility	*Data is within this article*

Value of the data

- The raw data containing drug-related chatter can be used to build prototype systems for a range of tasks in the domain of pharmacovigilance and toxicovigilance from social media, to assess user sentiments about the drugs, to estimate the effectiveness of distinct drugs and for other tasks important to the broader public health community.
- The distributed word representations were generated by varying multiple parameter combinations, thus ensuring that the different models capture different types of semantic information. Therefore, these models can be used for developing systems focused on mining knowledge associated with prescription medications.
- The n-gram language models capture sequential word occurrence probabilities and can be used for tasks such as text classification and text normalization.
- Python scripts are provided for downloading the tweets and for loading the two different sets of models.

1. Data

The data set consists of 267,215 Twitter posts, each of which contains at least one drug-related keyword. Two sets of language models accompany the raw data—the first is a set of models based on distributional semantics, which encapsulate semantic properties by representing word tokens as dense vectors, while the second set of models is based on n-gram sequences, capturing sequential patterns. All the data are available *via* our webpage, along with download/usage instructions: http://diego.asu.edu/Publications/Drugchatter.html. We will release more data and resources in the future via this link.

1.1. Characteristics of the data

The posts were collected over a four-month period—November, 2014 to February, 2015 and they contain over 250 mentions of unique drug-related keywords. The monthly distribution of the data is shown in Fig. 1. The figure suggests that the numbers of drug-related tweets collected are fairly consistent over the four months, with December seeing the highest number of drug-related posts, perhaps because it is holiday season. Fig. 1 also presents the frequencies of the top 10 drug-related keywords in the data sample we are releasing. ‘*adderall*’ is by far the most frequently found drug-related keyword.