Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics

Edidiong Okon, BSE,[a] Vishnutheja Rachakonda, BS,[a] Hyo Jung Hong, BA,[b] Chris Callison-Burch, PhD,[a] and Jules B. Lipoff, MD[c]
Philadelphia, Pennsylvania

Background: There is a lack of research studying patient-generated data on Reddit, one of the world's most popular forums with active users interested in dermatology. Techniques within natural language processing, a field of artificial intelligence, can analyze large amounts of text information and extract insights.

Objective: To apply natural language processing to Reddit comments about dermatology topics to assess for feasibility and potential for insights and engagement.

Methods: A software pipeline preprocessed Reddit comments from 2005 to 2017 from 7 popular dermatology-related subforums on Reddit, applied latent Dirichlet allocation, and used spectral clustering to establish cohesive themes and the frequency of word representation and grouped terms within these topics.

Results: We created a corpus of 176,000 comments and identified trends in patient engagement in spaces such as eczema and acne, among others, with a focus on homeopathic treatments and isotretinoin.

Limitations: Latent Dirichlet allocation is an unsupervised model, meaning there is no ground truth to which the model output can be compared. However, because these forums are anonymous, there seems little incentive for patients to be dishonest.

Conclusions: Reddit data has viability and utility for dermatologic research and engagement with the public, especially for common dermatology topics such as tanning, acne, and psoriasis. (J Am Acad Dermatol 2020;83:803-8.)

Key words: artificial intelligence; natural language processing; patient education; patient engagement; Reddit; social media.

Given ubiquitous Internet access, patients often research medical conditions themselves before seeking out professional expertise. Many patients also attempt self-treatments, informed by these web-based resources. Nearly two-thirds of adults in America use social media, a 10-fold increase over the previous decade.[1] Social media websites are commonly used as a source of medical information.[2] Because they are convenient, easily accessible, and often anonymous, social media may, compared with a face-to-face conversation with a physician, facilitate more open disclosure of symptoms experienced or remedies attempted. The rise of these electronic resources offers a unique opportunity to study how patients seek and consider dermatologic problems and

From the School of Engineering,[a] Perelman School of Medicine,[b] and the Department of Dermatology, Perelman School of Medicine,[c] University of Pennsylvania.

Funding sources: None.
Conflicts of interest: None declared.
IRB approval: Exempt by University of Pennsylvania IRB.
Accepted for publication July 3, 2019.
Reprints not available from the authors.

Correspondence to: Jules B. Lipoff, MD, Penn Medicine University City, 3737 Market St, Ste 1100, Philadelphia, PA 19104. E-mail: jules.lipoff@pennmedicine.upenn.edu.

Published online July 12, 2019.
0190-9622/$36.00

https://doi.org/10.1016/j.jaad.2019.07.014

treatments. Our goal was to build on previous work and evaluate the topics and advice being discussed on online forums such as Reddit.[3,4]

Natural language processing (NLP) is a subtype of artificial intelligence involving analysis of text for extracting insights.[5,6] These techniques have previously been applied to interpret narrative medical records.[7] In limited studies, NLP has also been applied to study medical insights from social media postings on Facebook and Twitter,[8] and in 1 study, Reddit.[9] Most recently in dermatology, NLP was demonstrated as an effective technique to quantify and characterize biopsy outcomes from pathology reports of melanocytic lesions.[10]

Past work in dermatology has considered the use of social media to find treatment options.[2,11] The online social media platform Reddit, which is the 18th most visited website in the world and sixth most in the United States,[12] is understudied relative to its popularity. Reddit has a forum-based interface, conducive to exchange of text and images. Anyone with Internet access can participate in user-generated subject-specific forums, dubbed *subreddits*, on particular topics. Users may post content in the form of ideas, questions, news article links, images, and videos. Thus, Reddit allows users to interact with geographically dispersed individuals who share similar experiences or topics of interest.

An exploratory study of Reddit identified its utility as a platform on which individuals exchange information about dermatologic topics.[3] Mining user-generated Reddit data has the potential to expand and refine providers' understanding of dermatologic conditions and therapies that patients undergo. However, given the large data volume and range of topics on Reddit, it is difficult for clinicians to gain actionable insights from the platform in its current format. To make the social media data more accessible and useful to clinicians, we used natural language processing techniques with the goal of creating a reference website, Reddermatology, to digest the dermatologic subset of raw data from Reddit and display key trends and metrics of crowd opinion, interest, and interaction. We pursued this goal by developing NLP-based computational methods and then analyzing their potential utility for clinical providers and researchers to enhance their knowledge of patients' experience of various dermatologic conditions, existing therapies, and novel alternative therapies.

CAPSULE SUMMARY

- The social media forum, Reddit, features extensive comments from patients about dermatology. Natural language processing is a technique that can extract insights.
- Data from social media forums may be extracted using artificial intelligence and leveraged to study patient engagement and interest in treatments and trends in dermatology.

METHODS

This study was deemed exempt by the University of Pennsylvania Institutional Review Board. The data for the study consisted of comments on forums of the freely available social media platform Reddit (http://www.reddit.com). Using the list of subreddits previously identified as relevant to dermatology,[3] a data set of 176,000 Reddit comments was gathered using a Google BigQuery (Google, Mountain View, CA) database infrastructure (Figure 1). The subreddit topics ranged across skin conditions (eg, psoriasis), body parts (eg, nails), medications (eg, isotretinoin), and popular consumer-oriented topics (eg, 30-plus skin care, skin care addictions). The database infrastructure was modeled off a 1.7-billion comment public resource created by Jason Baumgartner of pushshift.io. We queried the database using Structured Query Language and obtained 4 large text collections with the following metadata fields: name of subreddit, comment, comment author, and date of comment. Table I lists the dermatology-related subreddits used in this study.

Reddit comments from each of the 4 metadata fields were analyzed with NLP techniques. Initially, these comments were preprocessed into a text corpus. All comments were made lowercase, and punctuation was removed. The process of lemmatization removed the inflectional endings of words and returned their infinitive, or dictionary, definition. Stop words (eg, "a," "and," "or," "but," "how, etc) were removed from each Reddit comment. Because the data set contained multiple subreddits, the name of the subreddit from which the comment originated was appended to the end of each comment to improve the quality of the topics.

For each of the 4 metadata fields, the preprocessed text corpus was converted into a document-term matrix, which described the frequency with which terms appeared within a document. The document-term matrix is the input into the gensim implementation of latent Dirichlet allocation (LDA).[13] LDA is a generative probabilistic model often used for text corpora.[14] In LDA, which is a

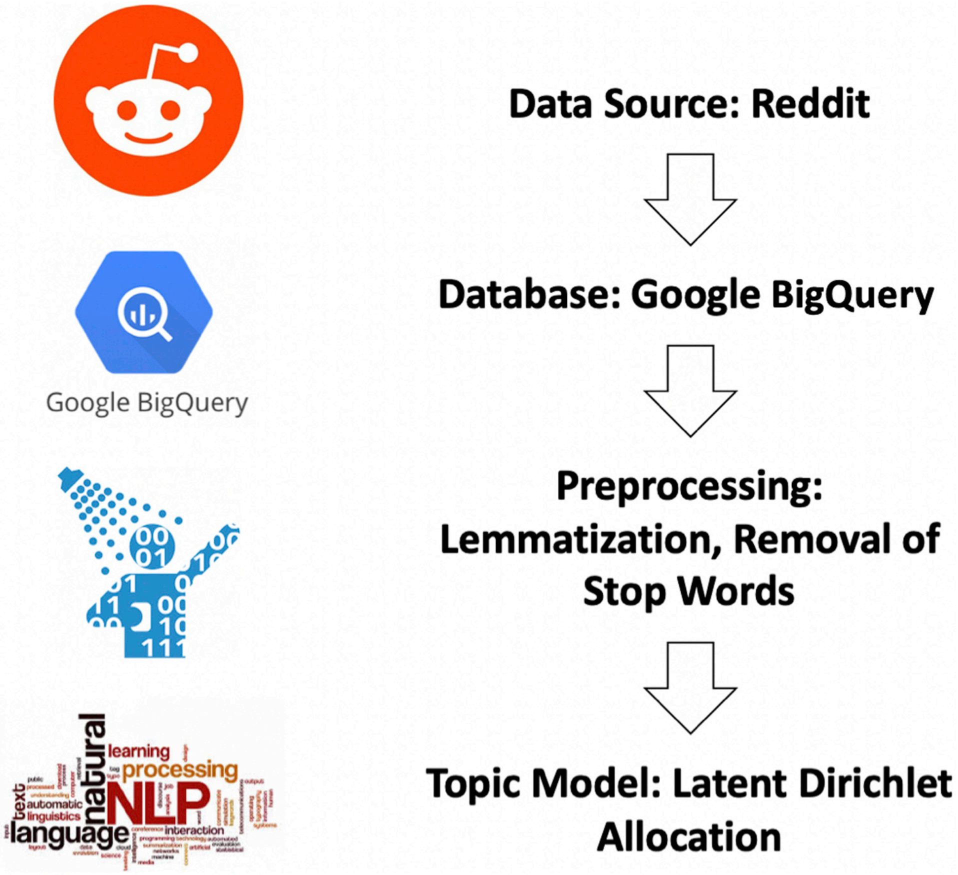
Data Source: Reddit
Google BigQuery
Database: Google BigQuery
Preprocessing: Lemmatization, Removal of Stop Words
Topic Model: Latent Dirichlet Allocation
natural
learning
processing
NLP
text
automatic
linguistics
language
interaction

Fig 1. The pipeline required data gathering from appropriate subreddits, formatting the data, and then training a topic model on the Reddit comment data. The user interface allows clinicians to qualitatively and quantitatively analyze the topic models. *NLP*, Natural language processing.

Table I. List of subreddits for comment aggregation with associated subscriber and comment counts as of August 2018

Subreddit	Total count of subscribers (No.)	Total count of comments (No.)
r/30PlusSkinCare	10,800	4579
r/Accutane*	8100	56,637
r/CompulsiveSkinPicking	17,000	23,093
r/Dermatology	5300	26,489
r/Nails	8200	5547
r/Psoriasis	8200	35,758
r/SkincareAddicts	25,900	23,993

*Roche, Cincinnati, Ohio.

3-level hierarchical Bayesian model, each collection, or set of documents (in this case, a set of comments), is modeled as a mixture of topics. Each topic is in turn modeled as a mixture over an underlying set of word probabilities, or word occurrences. Assuming a generative model for the collection of documents, LDA aims to efficiently approximate the Bayes parameters to find a set of topics that are likely to have generated the collection. The number of topics parameter, *k*, was manually attempted at escalating multiple group numbers (5, 10, etc) until a group level was found that produced a clear level of coherence for the topics, which occurred at 25.

The LDA topic model thus returned 25 topics. Within each topic was a list of words and each word's probability of appearing within the topic, as provided in Table II. For the terms within each topic, spectral clustering was run on the word2vec representations of the words with the goal of grouping terms within a topic for user readability, as shown in Table III.[15,16]

To facilitate visualization of the LDA topic model outputs, word cloud overviews were created for a subset of topics on common skin conditions (Fig 2): eczema, skin infection, psoriasis, and acne. The word clouds integrated input from both the spectral clustering and the topic probabilities. Each word cloud was visualized using the human-generated labels that spectral clustering enables. Next, each word in the cloud was sized and visualized based on

Table II. A subset of topics generated from our model*

	Human-generated descriptive topic label			
	"Eczema"	"Skin infection"	"Psoriasis"	"Acne"
Probability of each word appearing in the respective topic	Cream: 0.054 Treatment: 0.039 Steroid: 0.033 Topical: 0.024	Removed: 0.051 Infection: 0.041 Antibiotic: 0.030 Fungal: 0.018	Psoriasis: 0.095 Pain: 0.025 Humira‡: 0.017 Medication: 0.014	Accutane†: 0.091 Month: 0.080 Acne: 0.063 Course: 0.026

*Each term is associated with a topic with the displayed probability of occurring within in the respective model-derived topic. The top element of each column is the manually designated name of the topic. Among the total of 25 topics produced by the latent Dirichlet allocation model, the 4 most clinically pertinent topics are included.

†Roche, Cincinnati, Ohio.

‡AbbVie, Lake Bluff, Illinois.

Table III. A sample of the clustering output for the topic "itching"*

"Eczema"	"Skin infection"
['ointment,' 'itching,' 'hydrocortisone,' 'patch,' 'itch']	['pressure,' 'route,' 'middle']
['help,' 'scratch,' 'worse,' 'lot,' 'luck']	['sore,' 'headache,' 'lesion,' 'inflammation,' 'cyst']
['clobetasol,' 'effective,' 'apply,' 'option']	['chapstick,' 'cold,' 'chapped,' 'vinegar,' 'swollen']
['burn,' 'prescribed']	['rare,' 'common,' 'benign']
['treatment,' 'therapy,' 'treat,' 'biologic']	['yea,' 'word,' 'bet,' 'mention,' 'swears']
['spot,' 'time,' 'summer,' 'winter']	['apple,' 'cider,' 'turmeric']
['steroid,' 'tanning,' 'dermatologist,' 'prescription']	['anti,' 'eyebrow,' 'jawline']
['eczema,' 'spread,' 'outbreak']	['doctor,' 'science,' 'rheumatologist']
['cream,' 'lotion,' 'otc,' 'counter,' 'hand']	['meet,' 'appreciated,' 'agreed']
['topical,' 'skin,' 'cleansing,' 'primer']	['azelaic,' 'tane,' '120', 'vision']
['light,' 'sun,' 'sunlight']	['permanent,' 'wax,' 'adult']
['strong,' 'thin']	['infection,' 'antibiotic,' 'oral,' 'herpes,' 'staph']
['bed,' 'don't,' 'i've,' 'uvb,' 'mine']	['fungal,' 'cell,' 'bacteria']
	['removed,' 'disappeared']

*Each topic was derived from the unsupervised latent Dirichlet allocation model. The spectral clustering algorithm sorted words in a topic within the 25 topics generated by the model.

the percent probability of the word relative to others in the cloud.

Although the LDA model's values were useful to understand the most important words, they did not help generate order or give a thematic understanding of the discussions occurring on forums. For each of the 25 topics generated by the LDA model, a spectral clustering algorithm was applied to generate order to the probability values generated by the LDA models. In our experiments, the spectral clustering algorithm used the word2vec model of the words in each topic was applied. This algorithm thus produced clusters of words within each topic, allowing for readable interpretation of results generated by LDA model segregation. From the spectral clustering results, human interpretation was used to assign labels to the topics at a high level. To check the utility of the LDA model for dermatologic care providers, we performed queries and analyzed outputs for common dermatologic conditions: eczema, skin infection, psoriasis, and acne.

RESULTS

Comment aggregation and corpus creation

From the topics described in Dellavalle et al,[3] 20 subreddits related to dermatology were identified. The 7 dermatology-related subreddits with the greatest comment density from January 2005 to December 2017 were chosen for analysis. This number of subreddits was manually selected to focus study on the most substantial and recent engagement, meaning both high number of comments and recentness of activity. The total count of members and comments per subreddit are summarized in Table I. The most discussed subreddit was r/Accutane, followed by r/dermatology, and then r/skincareaddicts. The number of subscribers and comments on each subreddit varied widely, which is a function of organic adoption by users of a particular subreddit. Because the comment density of each subreddit varied, the entire comments set was aggregated into a larger corpus, with the origin subreddit appended to each comment.

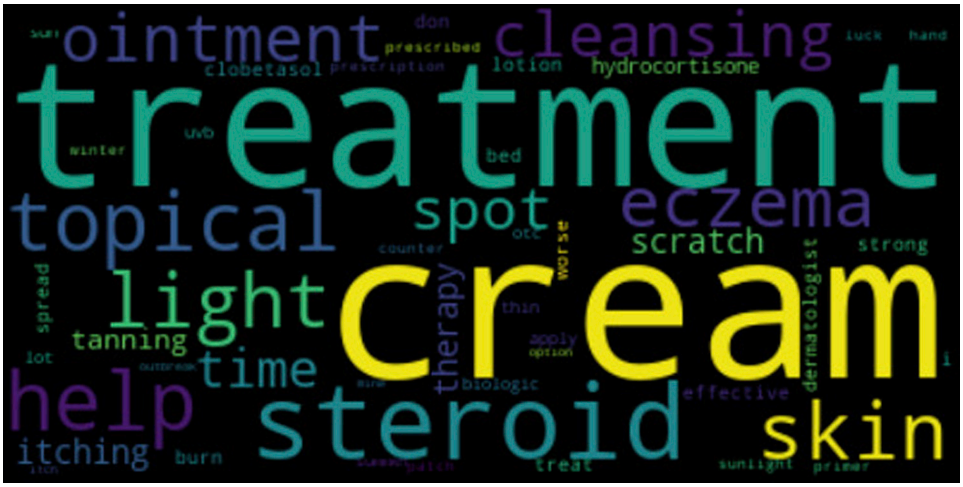

Fig 2. Word cloud overviews of an example topic: eczema. The size of each word corresponds to the frequency of occurrence in the topic.

Topic generation: Understanding frequency

A large corpus of comments was generated from the list of subreddits in Table I; 25 topics were produced by the LDA model. As described in the Methods, each of these topics is a collection of words. The entries in each column for each topic are the LDA-generated words associated with the topic, along with their corresponding probabilities.

Spectral clustering and topic labeling: Applying context

After applying a spectral clustering algorithm, human interpretation, with verification by a board-certified dermatologist (author J.B.L.), was used to assign high-level labels to the spectral clustering output. For example, a human reading the results of the spectral clustering algorithm in the "eczema" topic was able to identify that all of the words and clusters of words were related to eczema. Within the human-labeled topic, the specific clusters of words were mapped in relation to the overall topic. For example, in the "skin infection" topic (a human label), the cluster of ['apple,' 'cider,' 'turmeric'] clearly can be summarized as homeopathic treatments. This introduction of spectral clustering ordering allows for context to complement the probability values initially generated by the LDA model, which on its own focuses on frequency.

Word clouds

The word clouds suggested that discussions of Reddit users on the topic of eczema focused the most on treatment (Fig 2), and they were most likely to discuss creams, such as steroids, and other topical therapeutic options. In the skin infection word cloud, the words "fungal" and "antibiotic" were highly visible, indicating their relevance to the topic. In the psoriasis topic, analysis highlighted a number of different treatments and associated medicines such as "Humira" (AbbVie, Lake Bluff, IL).

DISCUSSION

Our data confirm that Reddit is a data resource with potential insights into patients and the medical treatments they are considering. First, doctors may learn about therapeutic options that patients might have tried or read about but are reluctant to bring up in a clinic visit. Second, use of such a forum's data may allow doctors to learn about treatments under discussion in the community (eg, tea tree oil and apple cider vinegar) that may prompt rigorous clinical study to investigate novel therapies, for example, as happened with naltrexone for Hailey-Hailey disease.[17]

Our results demonstrate that Reddit specifically is an effective source of user-generated information on various dermatology-related issues. Although these data did not highlight any one disease for physicians to follow-up with formal study, we did identify many trends in patient engagement. For example, the high number of subscribers to the Accutane subreddit demonstrated extreme interest about dosing and administration of the medicine. Similarly, using the unsupervised LDA model, we identified homeopathic treatments that patients discussed, which may be less recognized by physicians. For example, "dish soap" was found to occur together with other treatment related words, such as "apply." Through direct examination of the comments and investigation into both Reddit and other forums, we found that dish soap, specifically Dawn (Procter & Gamble, Cincinnati, OH), is often discussed and recommended by users as a treatment for acne.

This study is limited primarily by the inability to validate information posted on online forums. LDA is an unsupervised model, meaning there is no ground truth to which the model output can be compared. As a result, the quality of the outputs must be analyzed qualitatively. Whether anonymous patient comments accurately reflect treatments tried and reported effects is difficult to determine. However, because these forums are anonymous, there seems little incentive for patients to be dishonest. Further, there is inherent subjectivity in assigning human labels for these model-generated topics. Lastly, Reddit users are known to skew younger and more male than the general population, which could bias results.[18]

Social media-based topic model implementation may prove extremely useful for dermatology. It may represent an opportunity for physicians to correct misinformation and better connect patients with reliable, vetted clinical information. For instance, one possible future direction is to organize the LDA and clustering output into a website for clinicians. Thus, this new information from Reddit can help clinical professionals better understand how patients could be thinking about their conditions and their treatment options. These data can also inform physicians to offer proactive advice, for instance, advising against certain risky homeopathic treatments that are commonly shared on Reddit forums. Fundamentally, this new information can both enhance physicians' knowledge of patient behavior and then present an opportunity to act upon it. Beyond dermatology, this type of data mining can be expanded to leverage Reddit comments relevant to other fields of medicine to improve understanding of patient concerns and thoughts and to explore possible novel areas for clinical study.

CONCLUSION

Our study demonstrates Reddit data has viability and utility for dermatologic research and engagement with the public, especially for common dermatology topics such as tanning, acne, and psoriasis. Data from anonymous social media forums such as Reddit may be leveraged to study patient engagement and interest in treatments and trends in dermatology.

We acknowledge Robert P. Dellavalle, MD, PhD, MSPH, for study design suggestions and comments on the manuscript.

REFERENCES

1. Perrin A. *Social media usage, 2005-2015*. Pew Research Center; 2015. http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/. Accessed December 12, 2018.
2. Savas JA, Huang KE, Tuchayi SM, Feldman SR. Understanding the influence of social media in medicine: lesson learned from Facebook. *Dermatol Online J*. 2014;20(9).
3. Buntinx-Krieg T, Caravaglio J, Domozych R, Dellavalle R. Dermatology on Reddit: elucidating trends in dermatologic communications on the world wide web. *Dermatol Online J*. 2017;23(7).
4. Vance K, Howe W, Dellavalle RP. Social internet sites as a source of public health information. *Dermatol Clin*. 2009;27(2):133-136.
5. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551.
6. Friedman C, Hripcsak G. Natural language processing and its future in medicine. *Acad Med*. 1999;74(8):890-895.
7. Hripcsak G, Friedman C, Alderson PO, et al. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*. 1995;122(9):681-688.
8. Sarker A, Belousov M, Friedrichs J. 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. *J Am Med Inform Assoc*. 2018;25(10):1274-1283.
9. Blumenthal KG, Topaz M, Zhou L, et al. Mining social media data to assess the risk of skin and soft tissue infections from allergen immunotherapy. *J Allergy Clin Immunol*. 2019;144(1):129-134.
10. Lott JP, Boudreau DM, Barnhill RL, et al. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. *JAMA Dermatol*. 2018;154(1):24-29.
11. Menzies S, Daly S, McKenna D. Social media and psoriasis treatment: what are people saying on Twitter? *Br J Dermatol*. 2019;180(6):1527-1528.
12. Alexa. The top 500 sites on the web. Alexa.com. http://alexa.com/topsites/. Accessed January 28, 2020.
13. Blei D, Ng A, Jordan M. *Advances In Neural Information Processing Systems 14*. Cambridge: MIT Press; 2002:601-608.
14. Rehurek R, Sojka P. Software framework for topic modeling with large corpora. In: Proceedings of the LREC 2010 Workshop On New Challenges for NLP Frameworks, Valletta, Malta, May 22, 2010.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
16. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space; 2013. https://arxiv.org/abs/1301.3781. Accessed June 27, 2019.
17. Albers L, Arbiser J, Feldman R. Treatment of Hailey-Hailey disease with low-dose naltrexone. *JAMA Dermatol*. 2017;153(10):1018.
18. Sattelberg W. The Demographics of Reddit: who Uses the Site? TechJunkie; 2019. https://www.techjunkie.com/demographics-reddit/. Accessed June 27, 2019.