*Dieter Erben and Chang Liu*

*dev241 cl4855*

*New York University*

*Word Count: 2986*

*a dive into disaster relief news coverage*

DS-GA 3001: Text as Data

# Introduction

Large scale evacuation planning tools are becoming increasingly important due to the escalation in the frequency of natural disasters. The need to have a ready to use evacuation preparation tool that can be applicable anywhere in the world could save lives of many. Therefore, we want to learn about disaster related text data to get a better understanding of it. In addition, disaster relief after evacuation is essential to stabilize citizens and help the economy again stabilize after the impact. This research is conducted to understand disaster relief by learning disaster related text data. More specifically, we started with a disaster relief terms dataset and built a dictionary. Then we proceeded to scrape news data by filtered terms in this dictionary. Sentiment scores of different topics were also calculated by dictionary method to help us get a better understanding of the attitude towards disasters from news media. Lastly, we ran the LDA model to classify them into different topics and compare them with the original label topics of data.

# Literature

According to a study done by Thomas Eisensee and David Strömberg on the influence of mass media on US Government response, news coverage can have an enormous impact on disaster relief decisions[[1]](#footnote-1). Additionally, this news coverage can vary according to competition with other news, so if another major event is happening, disaster news coverage can decrease. Since this coverage translates directly into action, we were interested in learning how disaster relief is covered throughout the news. Brian Miles and Stephanie Morse argue that the media also plays a role in creating different perceptions of risks in the public’s minds when it comes to natural disasters[[2]](#footnote-2), further giving us evidence that news coverage plays a crucial role in disaster relief. After exploring the data, we used two methods to predict a label and compare it to the actual label.

First, we decided to look into analyzing a news article using defined dictionaries. We wanted to categorize our different news articles on how likely they were members of a certain class (topic) based on a pre-determined list of words.[[3]](#footnote-3) We used Young and Soroka’s idea of measuring sentiment with a dictionary-based approach, looking at the token frequency of a defined dictionary[[4]](#footnote-4). Our first implementation was very similar to what is described in this paper, using a dictionary of positive words and another one of negative words and measuring the total number of words in each dictionary present in the article, giving us a sentiment score. However, our second implementation used many dictionaries, and instead of having a positive or negative sentiment, we were interested in how much of a dictionary-based topic was present in each news article. While other methods have been implemented to measure emotion in news articles[[5]](#footnote-5), we decided to focus in dictionary based sentiment analysis and topic prediction.

Second, we looked into topic modeling with LDA, in order to measure how different articles were covering different topics. Our main goal was to understand the distribution of topics in our news articles and the probability of each article belonging to a specific topic[[6]](#footnote-6). Additionally, we wanted to use LDA as a predictive model to label our news articles, and then compare with the true labels. This was inspired on McAuliffe and Blei’s sLDA (supervised Latent Dirichlet Allocation)[[7]](#footnote-7) and other supervised topic models such as FLDA (Frequency-LDA)[[8]](#footnote-8), where they looked at the distribution of labels, dependencies among them and predicting response values for new documents. Later in this paper we will discuss our methodology for our LDA based predictions.

Lastly, we looked into similar problems being done with different kind of data. Clustering methods have been used for news articles but focusing on the retrieval of these articles and named entity extraction[[9]](#footnote-9), or identifying emerging topics of discussion on online discourse[[10]](#footnote-10) rather than analyzing their content or focusing on disaster relief like we are. Emotion analysis at the sentence level with K-NN[[11]](#footnote-11), prediction from the reader’s perspective as a multi-label classification problem[[12]](#footnote-12), and comparing Twitter and traditional media with topic modeling[[13]](#footnote-13) have been studied within the news articles domain, but not focusing on our disaster relief topic or on topic labeling, motivating us to explore this new subject.

# Building the disaster relief dictionary

While there where many data sources for news articles, we could not find any source for disaster relief news articles. Many of the sources were not even labeled, so we decided to get our own data. In order to do this, we started by looking into how disaster relief is defined and the keywords that relate to disaster relief. We got our first dataset at Figure Eight, which is a machine learning and artificial intelligence company that via crowdsourcing and algorithms, transcribes or annotates data. There, we found a dataset called “Relevance of terms to disaster relief topics”[[14]](#footnote-14), which had many different terms and phrases relevant to disaster relief. The data was grouped by 13 topics and many terms for each topic. A sample of the dataset can be found in Appendix A.

We used the relevance column as a confidence measure on the relevance of this term to the topic, and based on this number and our intuition, we selected 5 terms for each topic. These 65 terms defined our vocabulary for disaster relief, which we used for the next steps of our data gathering process. Once we defined these terms, our goal was to get news articles related to each of these terms, and in order to do this, we decided to scrape news articles for each keyword using Bing’s API.

# Scraping

We got our dataset by scraping data with disaster related keywords. We used Bing News API to scrape 100 links of news text sources by each keyword. Bing is a web search engine owned by Microsoft and their limited version open data is free for developers and provided by Microsoft Azure. Once we got links of news sources, we proceeded the scraping work with beautifulsoup to filter irrelevant information out like images but keep only text data from website.

# Preprocessing

After scraping the news articles for each topic, we started the preprocessing task. The first step here was to get rid of the failed scraping attempts. Several scrapes did not work as we intended due to API issues or the content of the website was not what we were looking for. Therefore, we dropped all news articles with less than 900 characters to get rid of these issues. Each article was further processed by removing non-alphanumeric characters, numbers, punctuation and converting all words into lowercase. The resulting text, with its corresponding keyword and topic, were used to create our main dataset. This was used in the models we will discuss below and to create document term matrices and more.

# Hypotheses

With the dictionary and LDA model, we will tag new labels on each text. We expected that the dictionary and LDA models to be able to label the text data. The null hypothesis is that new labels and original labels are significantly different and alternative hypothesis is they are equivalent. We used an F test because it compares 2 sample variances to see if they share same quantity of variance to tell whether they are significantly different or not. We ran F tests on original labels data between the dictionary labels and LDA labels outcome respectively to see how p value and F scores performed. We predict that our label generating methods will be effective when evaluated with this statistic.

# Methods and Results

## Dictionary

For different topics in disaster relief dataset, there are terms which can represent those topics. We built a dictionary for disaster topics using R’s dictionary function. With the dictionary we created a document feature matrix which indicated the frequency of terms that appeared in each text. This way, we got the frequency of terms distribution of different topics for each text and we chose the topic with the max value to label the text. This dictionary enabled us to predict any disaster related text that we can encounter in the future.

## Sentiment

It’s also very important to understand the sentiment of news in terms of different disaster topics. We used a large dictionary provided by WordNet which includes positive and negative words respectively. The final sentiment score for each text is calculated from the positive score subtracting the negative score. After we got the sentiment scores for each news article, we aggregated the dataset by topic and calculated the average sentiment score for each of them. It turned out the sentiment scores for all the topics were negative. This makes sense because topics are all disaster related. Furthermore, shelter related articles were the most negative and came along with the extreme violence and terrorism, infrastructure and facilities. The least negative one was intervention. This indicated the strong negativity from news media towards shelter and terrorism..

## LDA

With Latent Dirichlet Allocation[[15]](#footnote-15), we were interested in two things: 1) the proportion of topics found in different news articles, and 2) the creation of topics based on these different news articles. The first step in this model was to create a document term matrix for our disaster relief news articles corpus. For the preprocessing, we decided to stem words and remove stop words. We stemmed words in order to capture related words in the same stemmed word, allowing us to reduce the total number of words and get our topics to include different words rather than dividing the importance of a word because of its variations. After stemming, we removed stop words to capture better the essence of each topic and not confuse the model with common words for each topic such as stop words. However, after creating this data frame, we had more than 89,000 features. The number of features here made it difficult to run different models and after inspecting the data, we realized many of these features were words that appeared only in some documents. Therefore, we decided to trim the document feature matrix to only include words with a minimum frequency of 20, ensuring us it relates to either many topics or is very important for a specific topic. Once this subset of the document term matrix was created, we looked into the number of topics we wanted to have.

While we were considering creating k=13 topics for our LDA modeling, we realized that many topics were being grouped into the same topic with LDA. Therefore, we decided to instead look for the optimal number of topics by calling the FindTopicNumber[[16]](#footnote-16) function of LDA Tuning, which calculates different metrics to find this k. To find the number of topics, we decided to look at two metrics: Griffiths[[17]](#footnote-17) and CaoJuan[[18]](#footnote-18) to find the best number of topics and Gibbs[[19]](#footnote-19) as our sampling method. As seen in Appendix B, we decided to go with 25 topics. With 25 topics, we are still optimizing both metrics, yet also limiting the number of topics in order to not get too specific and lose the interpretation of the topics. Since we originally had 13 topics from our data, we thought 25 was a good number to use, supported by the graph in the appendix. At 25 topics, there is only a marginal gain when increasing topics. After fitting our LDA method, again with Gibbs sampling, we got a log likelihood of -13.448.664.

The first step after fitting our LDA model was to find the most common LDA topic corresponding to each of our original topics (labels). To do this, we found the most prevalent LDA topic in each article, based on the gamma matrix of the LDA model: P(topic|token). We then grouped the articles on their original topic label, and counted the frequency of each LDA topic by topic label. Lastly, we got the most frequent LDA topic for each topic label, and therefore getting our equivalent of the original 13 topic labels in terms of LDA topics. With this, we now looked at the most common words by topic. In order to do this, we focused on the beta of the LDA model, the Dirichlet prior for tokens over topics. By grouping over LDA topic, we got the top 10 words based on their beta values and gave us the most frequent tokens for each LDA topic. This, with our previous connection to the original topic labels, allowed us to generate the graph in Appendix C, which shows us the most common tokens for each LDA topic matching our original topic labels. By looking at the gamma matrix again, and the mapping we did of matching LDA topics to topic labels, we generated a prediction within our topic labels for each article. This prediction, giving us one of the 13 original topics, was used as a method of label creation for unlabeled news articles.

## Results

Using an F test, we got an F1 score between original labels and LDA labels of 0.781, while the F1 score between the original and dictionary labels is 0.468. Under 95% confidence interval, we will reject the null hypothesis when F is more than 0.175 and less than 5.12. Both of the F score of different labels laid in the interval so we can reject the null hypothesis, which is the outcome is significantly different from the original input. Therefore, we can say that both dictionary and LDA models can satisfyingly predict the outcome based on the F test. Our confusion matrices, seen in Appendix D, had accuracies of 78.6% for dictionary prediction and 53.9% for LDA model. This indicates that the dictionary prediction is slightly better than the LDA model. This could be because the dictionary is built with specific disaster relief terms that relate directly to the topic, while the LDA model tried to predict by words distribution which is much more vague. The simplistic approach of the dictionary probably performed better since its terms are focusing only on one topic, while the LDA can capture many different topics in an article and therefore predict the wrong label. For example, this could happen when a water scarcity article focuses on the infrastructure, being labeled as *infrastructure* by the LDA model but being a *water* article.

# Discussion

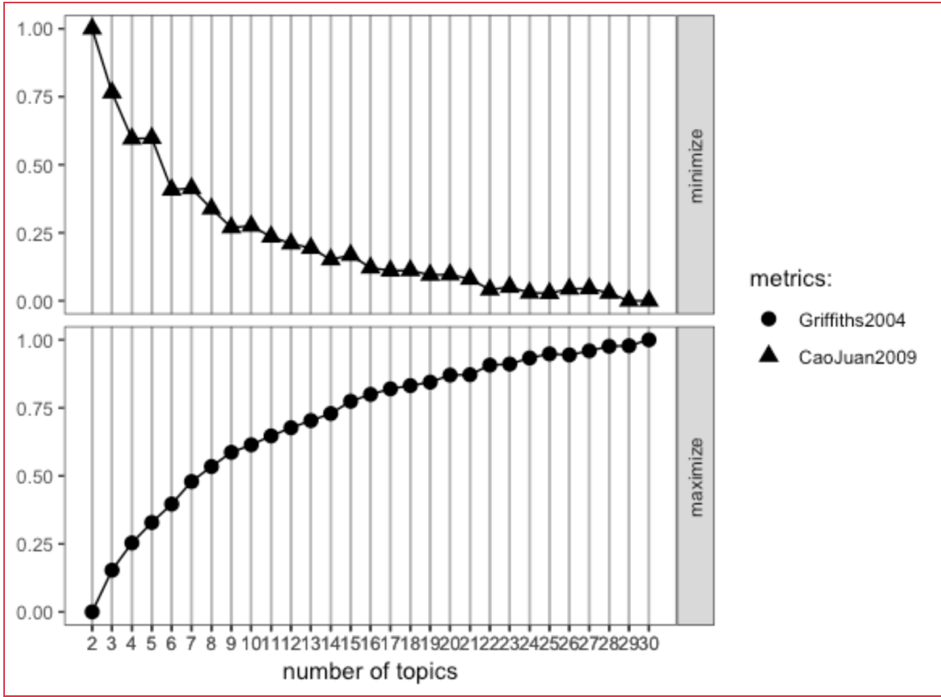
One of the limitations of our modeling was the source of our data. Our scraping method used the keywords we defined to get the top news articles in Bing for that specific API call at that time. If we were to run our scraping script again in a couple of months, the most relevant results would definitely change. Therefore, our results are very time sensitive, and to get a better understanding of disaster relief in general, we would have to scrape data for an extended period of time. This also varies depending on the topic, as certain topics tend to change a lot through time while others use the same vocabulary consistently.

In this paper, we combined our interest in disaster relief topics, with the methodology being used right now with news articles analysis. Additionally, we explored methods in sentiment analysis, dictionary-based methods and LDA to dive into disaster relief news coverage from a new perspective. While we focused solely on disaster relief, this analysis can be done in a similar way with a different subject, scraping articles with the subject’s vocabulary, building a dataset, and performing the same analysis done above to understand the topics within it and the way it is being covered by the media. This allows our study to be applicable to many different fields that could benefit from the study of news coverage of their subject. Lastly, it would also be interesting to see the differences in different countries and languages, as well as ideologies by news source.

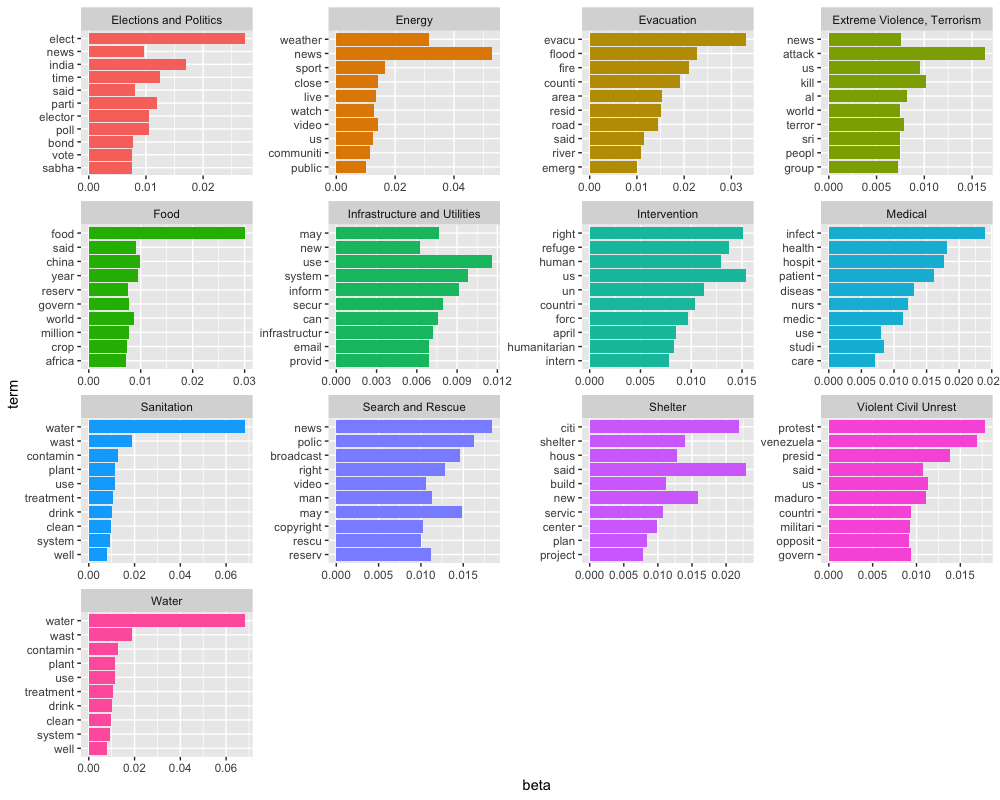
# Appendix A

|  |  |  |  |
| --- | --- | --- | --- |
| \_unit\_id | relevance | term | topic |
| 867851146 | 4.25 | AIR-STRIKE | Extreme Violence, Terrorism |
| 867851147 | 4.33 | DRINKING UNSAFE | Water |
| 867851149 | 4.33 | ELECTRICAL SUPPLIES | Energy |
| 867851156 | 4.67 | FIRED BULLETS | Violent Civil Unrest |
| 867851164 | 4.25 | TEMPORARY REFUGE | Shelter |

# Appendix B



# Appendix C



# Appendix D

## Dictionary Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| **1** | 33 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1 |
| **2** | 1 | 62 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 4 | 0 |
| **3** | 0 | 3 | 62 | 0 | 0 | 1 | 1 | 2 | 3 | 7 | 1 | 0 | 0 |
| **4** | 9 | 0 | 0 | 29 | 0 | 0 | 1 | 2 | 4 | 1 | 0 | 2 | 1 |
| **5** | 3 | 2 | 0 | 0 | 34 | 1 | 4 | 1 | 2 | 0 | 1 | 0 | 1 |
| **6** | 0 | 14 | 0 | 0 | 0 | 40 | 0 | 2 | 3 | 0 | 0 | 0 | 4 |
| **7** | 1 | 0 | 0 | 0 | 0 | 0 | 48 | 2 | 0 | 1 | 0 | 2 | 1 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 60 | 10 | 0 | 1 | 0 | 2 |
| **9** | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 3 | 45 | 0 | 0 | 0 | 27 |
| **10** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 13 | 0 | 0 | 1 |
| **11** | 1 | 1 | 1 | 0 | 0 | 3 | 0 | 1 | 2 | 0 | 24 | 1 | 0 |
| **12** | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 56 | 0 |
| **13** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 70 |

Overall Statistics

Accuracy : 0.7858

95% CI : (0.7543, 0.815)

No Information Rate : 0.1487

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7659

Mcnemar's Test P-Value : NA

## LDA Confusion Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| **1** | 26 | 0 | 0 | 1 | 1 | 0 | 3 | 0 | 0 | 1 | 0 | 5 | 0 |
| **2** | 1 | 22 | 10 | 13 | 13 | 2 | 10 | 2 | 2 | 4 | 7 | 3 | 0 |
| **3** | 2 | 16 | 47 | 2 | 0 | 0 | 4 | 3 | 0 | 2 | 9 | 8 | 1 |
| **4** | 2 | 1 | 0 | 29 | 2 | 0 | 2 | 1 | 0 | 12 | 0 | 0 | 0 |
| **5** | 2 | 0 | 0 | 1 | 36 | 1 | 3 | 3 | 0 | 1 | 3 | 0 | 0 |
| **6** | 7 | 7 | 2 | 0 | 1 | 30 | 4 | 1 | 0 | 4 | 0 | 0 | 6 |
| **7** | 6 | 3 | 0 | 6 | 0 | 3 | 28 | 1 | 0 | 1 | 0 | 11 | 0 |
| **8** | 1 | 3 | 0 | 0 | 1 | 4 | 2 | 56 | 0 | 2 | 3 | 2 | 3 |
| **9** | 1 | 0 | 1 | 0 | 6 | 2 | 1 | 0 | 22 | 2 | 13 | 1 | 17 |
| **10** | 0 | 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 8 | 0 | 2 | 1 |
| **11** | 4 | 1 | 3 | 3 | 1 | 0 | 3 | 0 | 0 | 1 | 15 | 2 | 0 |
| **12** | 2 | 2 | 0 | 7 | 1 | 2 | 3 | 0 | 10 | 0 | 3 | 32 | 1 |
| **13** | 3 | 9 | 4 | 4 | 2 | 1 | 4 | 1 | 0 | 1 | 0 | 2 | 43 |

Overall Statistics

Accuracy : 0.5389

95% CI : (0.502, 0.5754)

No Information Rate : 0.105

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4995

Mcnemar's Test P-Value : NA

1. Eisensee, T., & Stromberg, D. (2007). News Droughts, News Floods, and U. S. Disaster Relief. *The Quarterly Journal of Economics,122*(2), 693-728. doi:10.1162/qjec.122.2.693 [↑](#footnote-ref-1)
2. Miles, B., & Morse, S. (2007). The role of news media in natural disaster risk and recovery. *Ecological Economics,63*(2-3), 365-373. doi:10.1016/j.ecolecon.2006.08.007 [↑](#footnote-ref-2)
3. Spirling, Arthur. *Supervised Learning Techniques I*. Retrieved from New York University, Text as Data, February 26, 2019. [↑](#footnote-ref-3)
4. Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication,29*(2), 205-231. doi:10.1080/10584609.2012.671234 [↑](#footnote-ref-4)
5. Bhowmick, P. K., Basu, A., Mitra, P., & Prasad, A. (2009). Multi-label Text Classification Approach for Sentence Level News Emotion Analysis. *Lecture Notes in Computer Science Pattern Recognition and Machine Intelligence,*261-266. doi:10.1007/978-3-642-11164-8\_42 [↑](#footnote-ref-5)
6. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* *3*. Retrieved April 24, 2019, from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [↑](#footnote-ref-6)
7. Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121-128). [↑](#footnote-ref-7)
8. Li, X., Ouyang, J., & Zhou, X. (2015). Supervised topic models for multi-label classification. *Neurocomputing*, *149*, 811-819. [↑](#footnote-ref-8)
9. Toda, H., & Kataoka, R. (2005, May). A clustering method for news articles retrieval system. In *Special interest tracks and posters of the 14th international conference on World Wide Web* (pp. 988-989). ACM. [↑](#footnote-ref-9)
10. Kasiviswanathan, S. P., Melville, P., Banerjee, A., & Sindhwani, V. (2011, October). Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 745-754). ACM. [↑](#footnote-ref-10)
11. Bhowmick, P. K., Basu, A., Mitra, P., & Prasad, A. (2009, December). Multi-label text classification approach for sentence level news emotion analysis. In *International Conference on Pattern Recognition and Machine Intelligence* (pp. 261-266). Springer, Berlin, Heidelberg. [↑](#footnote-ref-11)
12. Ye, L., Xu, R. F., & Xu, J. (2012, July). Emotion prediction of news articles from reader's perspective based on multi-label classification. In 2012 International Conference on Machine Learning and Cybernetics (Vol. 5, pp. 2019-2024). IEEE. [↑](#footnote-ref-12)
13. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338-349). Springer, Berlin, Heidelberg. [↑](#footnote-ref-13)
14. https://www.figure-eight.com/data-for-everyone/ [↑](#footnote-ref-14)
15. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* *3*. Retrieved April 24, 2019, from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [↑](#footnote-ref-15)
16. <https://www.rdocumentation.org/packages/ldatuning/versions/0.2.0/topics/FindTopicsNumber> [↑](#footnote-ref-16)
17. T. Griffiths and M. Steyvers. Finding scientific topics. The National Academy of Sciences, 101:5228–5235, 2004. [↑](#footnote-ref-17)
18. Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing,* *72*(7-9), 1775-1781. doi:10.1016/j.neucom.2008.06.011 [↑](#footnote-ref-18)
19. Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web* (pp. 91-100). ACM. [↑](#footnote-ref-19)