ANALYSIS OF TOPIC MODELING WITH UNPOOLED AND POOLED TWEETS AND EXPLORATION OF TRENDS DURING COVID

Jaishree Ranganathan and Tsega Tsahai

Department of Computer Science,
Middle Tennessee State University, Murfreesboro, TN, USA

ABSTRACT

In this digital era, social media is an important tool for information dissemination. Twitter is a popular social media platform. Social media analytics helps make informed decisions based on people's needs and opinions. This information, when properly perceived provides valuable insights into different domains, such as public policymaking, marketing, sales, and healthcare. Topic modeling is an unsupervised algorithm to discover a hidden pattern in text documents. In this study, we explore the Latent Dirichlet Allocation (LDA) topic model algorithm. We collected tweets with hashtags related to corona virus related discussions. This study compares regular LDA and LDA based on collapsed Gibbs sampling (LDAMallet) algorithms. The experiments use different data processing steps including trigrams, without trigrams, hashtags, and without hashtags. This study provides a comprehensive analysis of LDA for short text messages using un-pooled and pooled tweets. The results suggest that a pooling scheme using hashtags helps improve the topic inference results with a better coherence score.

KEYWORDS

COVID, Latent Dirichlet Allocation, Tweets, Topic Modeling

1. INTRODUCTION

Social media is changing the way how information disseminates around the world. The speed at which information spreads has increased dramatically. The traditional roles of microblogs, social networking sites, media sharing technologies are changing with the proliferation of data transmitting technologies [1]. According to global social media statistics [2], as of October 2021, there are around 4.55 billion social media users around the world. It is approximately equivalent to 57.6% of the global population. There is a constant increase in the annual growth of social media users.

Twitter is one of the world's top social media platforms. A tremendous amount of information is shared using such microblog platforms. The use of social media text for analysis of information spread in communities and the cascading effects can help understand how deeply such forums influence the social well-being, perceptions, beliefs, public health, and political decisions [3]. Based on the different demographic categories of Twitter users, the data provided by Twitter is a diverse source for researchers and policymakers [4] [5] [6]. Twitter data has been widely used in a variety of different research areas including disaster management, community analysis, network analysis, stock market prediction, recommender systems, health discussion, sports/entertainment, and politics [7].

DOI: 10.5121/ijcsea.2021.11601

Analyzing tweets over a period could provide meaningful insights into what concerns people about a particular event. Topic modeling provides a probabilistic framework to automatically search, understand and summarize large unstructured text documents. Since its inception, topic modeling has been applied in variety of research domains including but not limited to political science [8], literary studies [9], sociology [10], and healthcare [11] [12]. Topic modeling is also a popular approach used with Twitter data. For instance, studies focused on analyzing football news [13], online analytical processing [14], analysis of linguistic signal for depression detection [15], HPV vaccine discussions [16], pandemic [17] [18].

There are various algorithms or methods in the literature to extract topics from a collection of documents. These algorithms include Latent Dirichlet Allocation (LDA) [19], Non-Negative Matric Factorization (NMF) [20], Latent Semantic Analysis (LSA) [21], Parallel Latent Dirichlet Allocation (PLDA) [22], Pachinko Allocation Model (PAM) [23]. Latent Dirichlet Allocation (LDA) is one of the most common topic modeling algorithms. Tweets are a popular source of research in a plethora of categories mentioned above. There are several limitations associated with tweets. Tweets are limited in the number of characters, which also includes, hashtags, URLs, user mentions. This prevents the full potential of text analysis which is substantially different from the traditional text analysis in information retrieval. It is noted that many of the studies in literature focused on using the LDA for Twitter data topic modeling applied basic pre-processing of the data. Limited studies use tweet pooling techniques based on conversation [24], users [25] or authors [26] [27], hashtags [28], and burst scores [29].

In this work, we analyse the use of LDA for short text messages (tweets). We perform a series of experiments with parameter tuning pre-processing. In addition, we also explore the topic modeling approach using tweet pooling based on hashtags [28]. Our study provides a comprehensive analysis of LDA for short text messages using unpooled and pooled tweets. To the best of our knowledge, there is not much work that performs an extensive set of experiments and provides analysis based on multiple settings using LDA on covid related dataset. The rest of this paper is organized as follows section 2 - literature review, section 3 - data collection, section 4 - data pre-processing, section 5 - methods, section 6 - experiments and results, section 7 - results discussion, section 8 - limitations, section 9 - conclusion and future works and references.

2. Literature Review

In this section we establish the background in topic modeling algorithms and Twitter data for topic modeling.

2.1. Topic Modeling

Machine Learning and statistics research have developed techniques for finding patterns of words in a collection of documents using probabilistic models called Topic Models. Topic modeling is an unsupervised machine learning technique. Topic models scan a set of documents, detect words and phrase patterns within documents, and automatically cluster word groups and similar expressions that best categorize the documents. In today's world, there is an explosion of electronic documents. There is an ever-increasing demand for tools and techniques that automatically organize, search, browse or index a large collection of documents [30]. Topic models are popular in a multitude of domains such as social well-being, perceptions, beliefs, public health, political decisions [3], and marketing [31] [32] [33]. Especially topic models play a major role in the field of natural language processing [34].

There is an exponential growth in healthcare data. Pharmaceutical and biotech companies, health care providers seek the use of machine learning and natural language processing techniques to manage, regulate and derive useful insights in patient care and monitoring. Topic model algorithms are popular in the healthcare field. Several studies perform empirical analysis using electronic medical records [35], clinical notes [36], patient survey [12], gene expression data and microarray data [11].

2.2. Topic Modeling using Twitter Data

Social media serves as a communication tool across the world. It is used in a multitude of ways. For instance, to communicate the latest news, discussing events in real-time, sharing personal opinions on products, events, and many other associated factors. Twitter social media is a popular platform used by millions of people across the world to propagate real-time information during natural disasters, public events, and disease outbreaks. Since the outbreak of the COVID pandemic, social media is overwhelmed with a plethora of discussions worldwide. Several researchers collected tweets from the beginning of the year 2020 and throughout for analyses on tweet data. In this section, we cover the existing literature using the COVID-related tweet dataset for topic modeling.

2.2.1. Location Based Analysis

Some of the recent studies collected data for specific geography to analyze the tweets. For instance, [17] collects 68000 randomly filtered messages from Twitter using the streaming API during March and April of the year 2020. They use the LDA topic model to extract ten topics with a coherence score of 56%. The authors focused on identifying the most topical issues related to the COVID-19 pandemic discussed by South Africans. Authors in [37], collect both English (approx. 3,332, 565) and Portuguese (approx. 3,155,277) language tweets for the duration of four months (April - May) in 2020. The tweets were collected using hashtags relevant to the covid 19 pandemic. Using Geonames service the tweets are geotagged. The major focus of the study is sentiment analysis using the Sentence BERT model and comparison with other traditional machine learning models. The authors also use the collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM) for topic modeling.

Similarly [38], use approximately 319, 524 English language tweets collected (January to May 2020) for North American countries. They analyzed tweets using LDA and nonnegative factorization methods. The study concluded that LDA results were more distinct in categories. This work also used the aspect-based sentiment analysis. They compared the results using a timeline analysis of the outbreak. However, there is no specific coherence score or inter-topic distance. Authors [39], collect 777,869 English language tweets about COVID-19 nonpharmaceutical interventions in six countries.
They analyse the trends in public perceptions.

2.2.2. General Analysis

The study [40] collects approximately 150,000 tweets with keywords or hashtags related to COVID19 for four weeks. The author applies the Latent Dirichlet Allocation (LDA) model. The model extracts five topics and shows the inter-topic distance. Also, performs sentiment analysis (positive and negative sentiments) and emotions for top keywords from the data. Authors in [41] hydrate English language tweets collected for a specific duration (March 11, 2020, through January 31, 2021). These tweets are originally collected using 13 covid related keywords by [42]. Then the tweets are further filtered specifically for vaccine-related keywords. The study used 1,499,421 unique tweets and identify 16 topics using LDA topic modeling and the sentiments associated with vaccine-related discussion on Twitter. The optimal number of topics is chosen based on factors other than the coherence score. Similarly, [43] use 866,527 unique English language tweets collected between January 1, 2020, to April 30, 2020. Focus on analyzing the topics and themes during this period on Twitter using Latent Dirichlet Allocation (LDA) topic modeling approach.

Authors in [44], collect 107,990 tweets related to COVID-19 between December 13, 2019, and March 9, 2020. They analyze the data using topic modeling to identify and explore discussion topics over time. Apply latent Dirichlet allocation algorithm to identify the most common tweet topics (6 topics based on highest coherence score) as well as to categorize clusters and identify themes based on the keyword analysis. The coherence score is not reported explicitly. Authors [45] use Latent Dirichlet allocation (LDA) on tweets collected from January 23 to March 7, 2020.

The work discussed in the literature using topic model for COVID-related tweet data, use the approach of each tweet as a separate document. In this work, we perform a series of experiments with parameter tuning, pre-processing. In addition, we also explore the topic modeling approach using tweet pooling based on hashtags [28]. Our study provides a comprehensive analysis of LDA for short text messages using unpooled and pooled tweets. To the best of our knowledge, there is not much work that performs an extensive set of experiments and provides analysis based on multiple settings using LDA on such large COVID-related tweets.

3. DATA COLLECTION

In this study, we use the Tweepy search API for retrieving tweets. We collect English language tweets between August 31, 2021, to October 12, 2021. The tweets are collected using hashtags (shown in Table 1) related to COVID-19. We collected approximately 1 million instances in this process.

Table 1. Hashtags.

Hashtags List
stayhome, quarantine, lockdown, stayathome, socialdistancing, staysafe, washyourhands, disinfectant, handwashing, mask, ppe, covidvaccine, covidvaccination, vaccine, coronavirusvaccine, boostershots, boostershot, pfizerbooster, deltavariant, coviddeltavariant, SARSCoV2, muvariant, mu, gammavariant, gamma, delta, coronavirus, covid, COVID19, corona, pandemic, coronaviruspandemic

4. DATA PRE-PROCESSING

Pre-processing is the major step in any natural language processing task. The first step in pre-processing is to remove noise from the text data. In this step, we process each tweet text using Python's regular expression library to remove the following: URLs, HTML tags, emoticons represented as text. We then use the natural language toolkit (NLTK) library [46] to remove the stop words. We then removed tweets that contained five or less than five words. Most of the tweets tend to use informal language. So, the tweet text is then processed to update the contractions using Python's contractions library. Further, each tweet is processed for abbreviations and or slang words. All tweets are converted to lowercase to reduce dimensionality.

After the initial noise removal, we performed the following steps to normalize the tweet text. Python's Textblob library is utilized to lemmatize each word in the tweet text. Lemmatization is the process which returns the base or dictionary form of the word. Then convert words in the tweet to the base form using NLTK's port-stemmer library. After the pre-processing, we have 950,899 instances in the dataset.

5. METHODS

In this section we explain the models, additional data processing, tweet pooling for the experiments.

5.1. Topic Model Algorithms

There are various algorithms or methods in the literature to extract topics from a collection of documents. In this study, we use Latent Dirichlet Allocation (LDA) [19], one of the topic model algorithms to classify the text in documents into different topics. It helps discover hidden insights from the set of documents using Dirichlet distributions over the topics and the words. The following are the steps involved in LDA algorithm: first a vocabulary is generated based on the documents; then the algorithm produces a mix of topics for each document; next the algorithm determines mix of words for each topic by using relative frequency. We also explore genism LDA Mallet [47]. It is a Python genism wrapper for LDA based on MALLET [48] [47] a java topic modeling toolkit. LDA Mallet module uses collapsed Gibbs sampling [49] a Markov-chain Monte Carlo method from MALLET [48]. To determine the best topic models, we use the coherence score measure. This score is based on a sliding window that uses normalized pointwise mutual information (NPMI) and cosine similarity. We use the Python Gensim [50] for LDA models and Pythons LDAvis [51] for visualization of the results.

5.2. Additional Data Processing

In the data collection process, we use specific key hashtags related to COVID discussions on Twitter as shown in Table 1. However, there are other hashtags associated with each tweet. After each tweet is pre-processed, we only include the key hashtags (refer Table 1) as part of the final processed tweet. We conduct experiments without key hashtags and with key hashtags to see the performance of the topic model algorithms.

5.3. Tweet Pooling

In this study, we utilize hashtag-based pooling. The tweets are pooled based on the hashtags used for data collection. For example, if a tweet consisted of hashtags such as *#coronavirus, #covid, #COVID19, #corona, #pandemic, or #coronaviruspandemic,* it will be placed under group 3 (refer Table 2). If a tweet does not contain hashtags from Table 1, it is marked as an unknown group. Table 2 shows the hashtags along with the tweet distribution among the groups. We apply topic importance based on the tweet distribution and generalization of hashtags as factors while pooling the tweets. i.e., The unknown category is not assigned to any specific pool.

Table 2. Tweet Pooling Based on Hashtags.

Group	Hashtag	Distribution
0	deltavariant, coviddeltavariant, SARSCoV2, muvariant, mu, gammavariant, gamma, delta	31,032 (3.3%)
1	covidvaccine, covidvaccination, vaccine, coronavirusvaccine, boostershots, boostershot, pfizerbooster	101,231(10.6%)
2	stayhome, quarantine, lockdown, stayathome, socialdistancing, staysafe, washyourhands, disinfectant, handwashing, mask, ppe	74,029 (7.8%)
3	coronavirus, covid, COVID19, corona, pandemic, coronaviruspandemic	730,461(76.8%)
Unknown	-	14,146(1.5%)

6. EXPERIMENTS AND RESULTS

6.1. Experiment 1 – Using LDA Model and Pooled Tweets

Model 1 in Table 3, use tweets from all groups in Table 2 except the unknown group. The number of Instances is now 936,753. Then we concatenate tweets from each of the groups into an individual document. For example, all the tweets in the group 0 from Table 2 are concatenated to represent one document. As a result of this, we have four documents.

We use the Gensim tool to tokenize the documents and create bigrams and trigrams. Inputs to the LDA model are dictionary and corpus. Dictionary contains an id for each word and the corpus contain word ids mapped to the frequency of the words in a document. We choose the number of topics as 3 and 4.

Model 2 in Table 3 is very similar to model 1. It uses four documents to train the model. The only difference is that trigrams and bigrams were not added to these documents. In model 3 all the hashtags that were part of the tweets are removed. For this model, we included bigrams and trigrams. Model 4, like model 3 is trained without the bigrams and trigrams.

Table 3. Experiments Using LDA Model and Pooled Tweets.

Model	Description	Coherence Score	
		3 Topics	4 Topics
1	With Bigram/Trigram, with key Hashtags	0.511	0.396
2	No Bigram/Trigram, with key Hashtags	**0.561**	0.403
3	With Bigram/Trigram, no Hashtags	0.377	0.288
4	No Bigram/Trigram, no Hashtags	0.431	0.389

6.2. Experiment 2 – Using LDA Model and Unpooled Tweets

In this experiment, each tweet is considered an individual document. There are 950,899 documents. These tweets included the hashtags that were in the search query. The only difference between the two models is the inclusion and exclusion of bigrams and trigrams. The models are trained for 10, 15, 20, 25, and 30 topics. The coherence score for each of the models is shown in Table 4.

Table 4. Experiments Using LDA Model and Unpooled Tweets.

Model	**Description**	**Coherence Score**				
		10 Topics	15 Topics	20 Topics	25 Topics	30 Topics
1	With Bigram/Trigram, with key Hashtags	0.453	0.465	0.476	0.476	0.457
2	No Bigram/Trigram, with key Hashtags	0.468	0.467	0.454	0.423	0.430

6.3. Experiment 3 – Using LDAMallet Model and Pooled Tweets

In this experiment, LDA Mallet was used to train models. All the LDA Mallet models in this experiment is trained with the documents that are pooled based on hashtags. The difference is inclusion and exclusion of bigrams, trigrams, and hashtags. Results for the models with description is shown in Table 5.

Table 5. Experiments Using LDAMallet Model and Pooled Tweets.

Model	**Description**	**Coherence Score**	
		3 Topics	4 Topics
1	With Bigram/Trigram, with key Hashtags	0.357	0.372
2	No Bigram/Trigram, with key Hashtags	0.441	0.450
3	With Bigram/Trigram, no Hashtags	0.327	0.350
4	No Bigram/Trigram, no Hashtags	0.427	0.425

6.4. Experiment 4 – Using LDAMallet Model and Unpooled Tweets

Similar to the Experiment 2, the two models in this experiment are trained on each individual tweet as an input document. These tweets also included the hashtags that were in the search query (Table 1). The models differ based on inclusion and exclusion of bigrams, and trigrams and trained on 10, 15, 20, 25, and 30 topics.
Results for these models are depicted in Table 6.

Table 6. Experiments Using LDAMallet Model and Unpooled Tweets.

Model	**Description**	**Coherence Score**				
		10 Topics	15 Topics	20 Topics	25 Topics	30 Topics
1	With Bigram/Trigram, with key Hashtags	0.462	**0.501**	0.507	0.516	0.536
2	No Bigram/Trigram, with key Hashtags	0.484	0.482	0.516	0.521	0.529

7. RESULTS DISCUSSION

In this section we discuss the top representative models from section 6. Based on the coherence scores, the best result is achieved by model 2 in Table 3. The coherence score is 0.561. It is evident that including the hashtags in the documents yielded better coherent scores and topics compared to other models. Table 7. explains the inferred themes for each topic from the Figure 1. The distance between the bubbles in the graph represents how similar or different the topics are to each other. For example, topics 1 and 2 are closer to each other compared to topic 3. As we see topic 1 and 2 contains words related to covid and vaccines. However, topic 2 is more relevant to vaccines. So, topics 1 and 2 are more similar than topics 1 and 3 or 2 and 3. The most salient words in topic 3 are related to the variants.

We choose the optimum number of topics for the model as three, as this achieved the best result both in terms of better topic inference and coherence score. This outcome could be due to the distribution of tweet categories in the documents. We also compared the documents and the topics using topic dominance measure. The same is shown in Table 8.

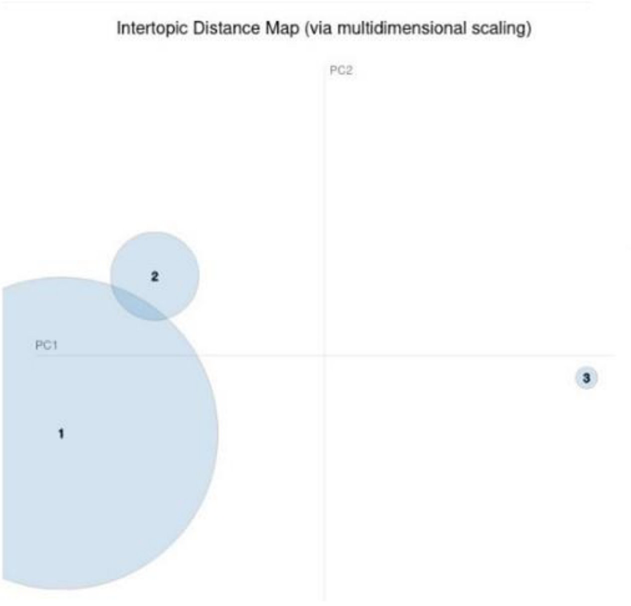

Figure 1. Intertopic Distance for Model 2 from Table 3.

Table 7. Topic Inference for Model 2 from Table 3.

Topic Number	Top Terms	Inferred Topic Theme
Topic 1	covid, vaccin, case, new, coronavirus	General Information - Covid
Topic 2	vaccine, dose, covid, vaccin, covidvaccine, age	Vaccine
Topic 3	deltavariant, delta, variant, muvariant, mu	Variants

We observe that document themes match the topics inferred. However, there are some deviations which we address as follows: The distribution of tweets in group 0 and group 2 is the lowest compared to the group 1 and 3 (refer Table 2). So, tweets in these documents are dominantly categorized to a general COVID theme. As most of the tweets contain the hashtags *#coronavirus, #covid, #COVID19, #corona, #pandemic, #coronaviruspandemic* etc. The topics inferred fall into categories of covid, vaccine and variants.

Table 8. Documents vs Topic Dominance for Model 2 from Table 3.

Document/Group	Dominant Topic 1	Dominant Topic 2
0	1 (covid)	3 (variant)
1	2 (vaccine)	1 (covid)
2	1 (covid)	-
3	1 (covid)	-

Among the other experiments, the coherence scores are better for the LDA Mallet model 1 with unpooled tweets shown in Table. 6. We choose the 15-topic model with a coherence score of 0.501. We could see the inter-topic distance is low and some of the topics are clustered and overlapped from Figure 2.

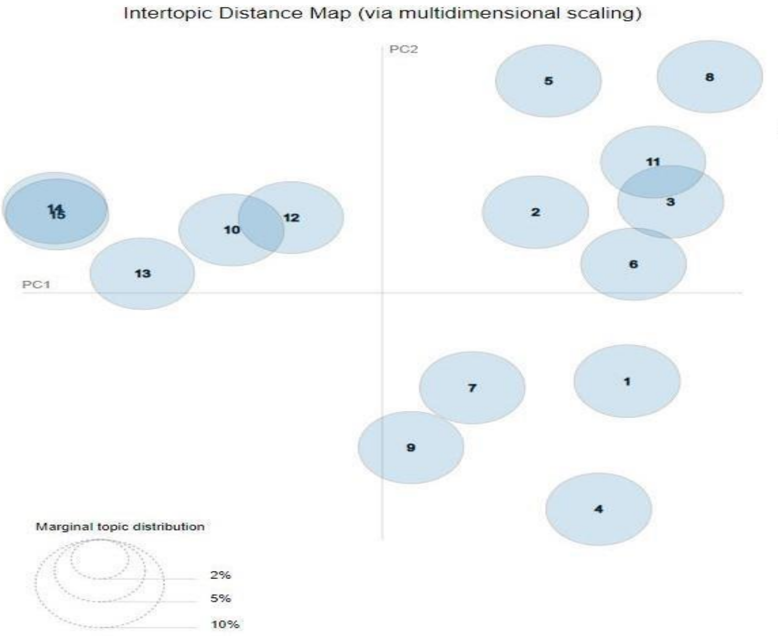

Figure 2. Intertopic Distance for Model 1 from Table 6.

However, it is found that some of the inferred topics give some specific insights. Table 9 shows the top inferred topics based on the result of this model. For instance, some of the representative themes are from topics 1, and 4 are about risks, prevention, and treatment, handling the covid pandemic situation; topics 12, 15 are about the covid vaccine and its administration.

Table 9. Representative Topics for Model 1 from Table 6.

Topic Number	Top Terms	Inferred Topic Theme
Topic 1	infect, viru, risk, delta, effect, studi, variant, show, prevent, sarscov, deltavariant	Covid risks, prevention, and treatment
Topic 4	countri, world, support, pandem, global, continu, crisi, meet, develop, address	About handling pandemic situation
Topic 12	vaccin, covid, dose, covidvaccine, fulli, shot, booster, age, receiv, jab	Vaccine
Topic 15	covid, vaccine, hospit, care, patient, dose, vax, administ, medic, unvaccine, doctor	Vaccine administration

This nature of the results could be due to the inherent complexity of short text tweets. Pooling the tweets together in documents based on specific criteria helps improve the coherence score and inference with distinct topics. However, from the results we infer that the distribution of data in each document plays an important role. Having a normalized set of data for each document could help detect detailed topics in the case of tweet pooling.

8. Limitations

One of the limitations is the hashtags used to retrieve tweets related to the covid 19 pandemic. More relevant tweets could have been missed in the data collection process if the tweets did not include the key hashtags in the Table 1. In the tweet pooling process, some of the tweets with hashtags related to a particular category which are not part of Table 2, could easily be pooled into a different document set. For instance, if a tweet had hashtags *#masks, #lockdowns* that would not match the category of group 2. The tweets retrieved are only a sample of the actual tweet inflow on Twitter.

9. Conclusions and Future Works

In this paper, we present an exploratory analysis of the results of topic modeling on tweets based on COVID related hashtags. We compare the results based on pooled and unpooled tweets using LDA and LDAMallet models. Our analysis demonstrates that the LDA topic model algorithm provides better topic inferences when operating over tweets pooled using hashtags. We evaluate our results using topic dominance measures to verify the mapping between document themes and the actual topic inferred. This work presents analysis of LDA topic modeling using un-pooled and pooled tweets. In this study we compare various models, based on parameter tuning, tweet pooling techniques and discuss the limitations. We plan to extend our work in the following aspects: firstly, use the preliminary topic inference results as a pipeline input to further process the data, by use of transfer learning; secondly, utilize big data to explore the performance and efficiency in data processing by use of cloud and distributed computing.

References

[1] Mayfield, Antony. “What is social media.” (2008).

[2] Datareportal (2021), “global social media stats”, retrieved from https://datareportal.com/socialmedia-users

[3] Vosoughi, Soroush, Deb Roy, and Sinan Aral. “The spread of true and false news online.” (2018) Science, vol. 359, no. 6380, pp. 1146-1151.

[4] Sarker, Abeed, Annika DeRoos, and Jeanmarie Perrone. “Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework.” (2020) Journal of the American Medical Informatics Association, vol. 27, no. 2, pp. 315-329.

[5] Hino, Airo, and Robert A. Fahey. “Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints.” (2019) International journal of information management, vol. 48, pp.175-184.

[6] Aladwani, Adel M. “Facilitators, characteristics, and impacts of Twitter use: Theoretical analysis and empirical illustration.” (2015) International Journal of Information Management, vol. 35, no. 1, pp. 15-25.

[7] Karami, Amir, Morgan Lundy, Frank Webb, and Yogesh K. Dwivedi. “Twitter and research: a systematic literature review through text mining.” (2020) IEEE Access, vol. 8, pp. 67698-67717.

[8] Grimmer, Justin. “A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.” (2010) Political Analysis, vol. 18, no. 1, pp. 1-35.

[9] Jockers, Matthew L., and David Mimno. “Significant themes in 19th-century literature.” (2013) Poetics, vol. 41, no. 6, pp. 750-769.

[10] DiMaggio, Paul, Manish Nag, and David Blei. “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding.” (2013) Poetics vol. 41, no. 6, pp. 570-606.

[11] Zhao, Weizhong, Wen Zou, and James J. Chen. "Topic modeling for cluster analysis of large biological and medical datasets." (2014) In BMC bioinformatics, vol. 15, no. 11, pp. 1-11. BioMed Central.

[12] Doing-Harris, Kristina, Danielle L. Mowery, Chrissy Daniels, Wendy W. Chapman, and Mike Conway. “Understanding patient satisfaction with received healthcare services: a natural language processing approach.” (2016) In AMIA annual symposium proceedings, vol. 2016, p. 524. American Medical Informatics Association.

[13] Hidayatullah, Ahmad Fathan, Elang Cergas Pembrani, Wisnu Kurniawan, Gilang Akbar, and Ridwan Pranata. “Twitter topic modeling on football news.” (2018) In 2018 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 467-471. IEEE.

[14] Yu, Dongjin, Dengwei Xu, Dongjing Wang, and Zhiyong Ni. (2019) “Hierarchical topic modeling of Twitter data for online analytical processing.” (2019) IEEE Access, vol. 7, pp. 12373-12385.

[15] Resnik, Philip, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. “Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter.” (2015) In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 99-107.

[16] Surian, Didi, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G. Dunn. “Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection.” (2016) Journal of medical Internet research, vol. 18, no. 8, pp. e6045.

[17] Mutanga, Murimo Bethel, and Abdultaofeek Abayomi. “Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach.” (2020) African Journal of Science, Technology, Innovation and Development, pp. 1-10.

[18] Prabhakar Kaila, Dr, and Dr AV Prasad. “Informational flow on Twitter–Corona virus outbreak–topic modelling approach.” (2020) International Journal of Advanced Research in Engineering and Technology (IJARET), vol. 11, no. 3.

[19] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation.” (2003) the Journal of machine Learning research, vol. 3, pp. 993-1022.

[20] Lee, Daniel D., and H. Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization.” (1999) Nature, vol. 401, no. 6755, pp. 788-791.

[21] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. “An introduction to latent semantic analysis.” (1998) Discourse processes, vol. 25, no. 2-3, pp. 259-284.

[22] Wang, Yi, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. “Plda: Parallel latent dirichlet allocation for large-scale applications.” (2009) In International Conference on Algorithmic Applications in Management, pp. 301-314. Springer, Berlin, Heidelberg.

[23] Li, Wei, and Andrew McCallum. “Pachinko allocation: DAG-structured mixture models of topic correlations.” (2006) In Proceedings of the 23rd international conference on Machine learning, pp. 577-584.

[24] Alvarez-Melis, David, and Martin Saveski. “Topic modeling in twitter: Aggregating tweets by conversations.” (2016) In Tenth international AAAI conference on web and social media.

[25] Hong, Liangjie, and Brian D. Davison. “Empirical study of topic modeling in twitter.” (2010) In Proceedings of the first workshop on social media analytics, pp. 80-88.

[26] Rosen-Zvi, Michal, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. “Learning author-topic models from text corpora.” (2010) ACM Transactions on Information Systems (TOIS), vol. 28, no. 1, pp.1-38.

[27] Weng, Jianshu, Ee-Peng Lim, Jing Jiang, and Qi He. “Twitterrank: finding topic-sensitive influential twitterers.” (2010) In Proceedings of the third ACM international conference on Web search and data mining, pp. 261-270.

[28] Mehrotra, Rishabh, Scott Sanner, Wray Buntine, and Lexing Xie. “Improving lda topic models for microblogs via tweet pooling and automatic labeling.” (2013) In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 889-892.

[29] Naaman, Mor, Hila Becker, and Luis Gravano. “Hip and trendy: Characterizing emerging trends on Twitter.” (2011) Journal of the American Society for Information Science and Technology, vol. 62, no. 5, pp. 902-918.

[30] Alghamdi, Rubayyi, and Khalid Alfalqi. “A survey of topic modeling in text mining.” (2015) Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 6, no. 1.

[31] Hu, Nan, Ting Zhang, Baojun Gao, and Indranil Bose. “What do hotel customers complain about? Text analysis using structural topic model.” (2019) Tourism Management, vol. 72, pp. 417-426.

[32] Park, Eunhye Olivia, Bongsug Kevin Chae, Junehee Kwon, and Woo-Hyuk Kim. “The effects of green restaurant attributes on customer satisfaction using the structural topic model on online customer reviews.” (2020) Sustainability vol. 12, no. 7, p. 2843.

[33] Rossetti, Marco, Fabio Stella, and Markus Zanker. “Analyzing user reviews in tourism with topic models.” (2016) Information Technology & Tourism, vol. 16, no. 1, pp. 5-21.

[34] Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey.” (2019) Multimedia Tools and Applications, vol. 78, no. 11, pp. 15169-15211.

[35] Huang, Zhengxing, Wei Dong, and Huilong Duan. “A probabilistic topic model for clinical risk stratification from electronic health records.” (2015) Journal of Biomedical Informatics, vol. 58, pp. 28-36.

[36] Chan, Katherine Redfield, Xinghua Lou, Theofanis Karaletsos, Christopher Crosbie, Stuart Gardos, David Artz, and Gunnar Rätsch. “An empirical analysis of topic modeling for mining cancer clinical notes.” (2013) In 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 56-63. IEEE.

[37] Garcia, Klaifer, and Lilian Berton. “Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA.” (2021) Applied Soft Computing, vol. 101, pp. 107057.

[38] Jang, Hyeju, Emily Rempel, David Roth, Giuseppe Carenini, and Naveed Zafar Janjua. “Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis.” (2021) Journal of medical Internet research, vol. 23, no. 2, p. e25431.

[39] Doogan, Caitlin, Wray Buntine, Henry Linger, and Samantha Brunt. “Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of twitter data.” (2020) Journal of medical Internet research, vol. 22, no. 9, p. e21419.

[40] Lee, Jae Hyun. “Understanding Public Attitudes Toward COVID-19 with Twitter.” (2021) In 2021 Systems and Information Engineering Design Symposium (SIEDS), pp. 1-6. IEEE.

[41] Lyu, Joanne Chen, Eileen Le Han, and Garving K. Luli. “COVID-19 vaccine–related discussion on Twitter: topic modeling and sentiment analysis.” (2021) Journal of medical Internet research, vol. 23, no. 6, p. e24435.

[42] Banda, Juan M., Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. “A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.” (2021) Epidemiologia, vol. 2, no. 3, pp. 315-324.

[43] Agarwal, Ankita, Preetham Salehundam, Swati Padhee, William L. Romine, and Tanvi Banerjee. “Leveraging Natural Language Processing to Mine Issues on Twitter During the COVID-19 Pandemic.” (2020) In 2020 IEEE International Conference on Big Data (Big Data), pp. 886-891. IEEE.

[44] Boon-Itt, Sakun, and Yukolpat Skunkan. “Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study.” (2020) JMIR Public Health and Surveillance, vol. 6, no. 4, p. e21978.

[45] Xue, Jia, Junxiang Chen, Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu. “Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter.” (2020) PloS one, vol. 15, no. 9, p. e0239441.

[46] Hardeniya, Nitin, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. “Natural language processing: python and NLTK.” (2016) Packt Publishing Ltd.

[47] Graham, Shawn, Scott Weingart, and Ian Milligan. “Getting started with topic modeling and MALLET.” (2012) The Editorial Board of the Programming Historian.

[48] McCallum, Andrew Kachites. “Mallet: A machine learning for language toolkit.” (2002) http://mallet. cs. umass. edu.

[49] Griffiths, Thomas L., and Mark Steyvers. “Finding scientific topics.” (2004) Proceedings of the National academy of sciences, vol. 101, no. suppl 1, pp. 5228-5235.

[50] Rehurek, Radim, and Petr Sojka. “Software framework for topic modelling with large corpora.” (2010) In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.

[51] Sievert, C., K. Shirley, and L. Davis. “A method for visualizing and interpreting topics.” (2014) In Proceedings of Workshop on Interactive Language Learning, Visualization, and Interfaces, Association for Computational Linguistics, pp. 63-70.