Tweetology of Learning Analytics

What does Twitter tell us about the trends and development of the field?

Khalil, Mohammad; Wong, Jacqueline; Er, Erkan; Heitmann, Martin; Belokrys, Gleb

DOI
10.1145/3506860.3506914

Publication date
2022

Document Version
Accepted author manuscript

Published in
LAK 2022 - Conference Proceedings

Tweetology of Learning Analytics: What does Twitter tell us about the trends and development of the field?

First Author's Name, Initials, and Last name

First author's affiliation, an Institution with a very long name, xxxx@gmail.com

Second Author's Name, Initials, and Last Name

Second author's affiliation, possibly the same institution, xxxx@gmail.com

Twitter is a very popular microblogging platform that has been actively used by scientific communities to exchange scientific information and to promote scholarly discussions. The present study aimed to leverage the tweet data to provide valuable insights into the development of the learning analytics field since its initial days. Descriptive analysis, geocoding analysis, and topic modeling were performed on over 1.6 million tweets related to learning analytics posted between 2010-2021. The descriptive analysis reveals an increasing popularity of the field on the Twittersphere in terms of number of users, twitter posts, and hashtags emergence. The topic modeling analysis uncovers new insights of the major topics in the field of learning analytics. Emergent themes in the field were identified, and the increasing (e.g., Artificial Intelligence) and decreasing (e.g., Education) trends were shared. Finally, the geocoding analysis indicates an increasing participation in the field from more diverse countries all around the world. Further findings are discussed in the paper.

CCS CONCEPTS • Human-centered computing → Social media • Applied computing→ Document management and text processing • Information systems→ Information retrieval→ Document representation→ Content analysis and feature selection

Additional Keywords and Phrases: Twitter, Twitter analysis, topic modeling, learning analytics, geospatial analysis

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA.

1 INTRODUCTION

Twitter is one of the most used social networking platforms worldwide, with more than 350 million monthly active users, generating 500 million tweets every day [Ahlgren 2020]. Twitter is a microblogging service, where users are allowed to post short messages (i.e., tweets) that are no longer than 280 characters (which has doubled from 140 in 2017). Microblogging is considered a new form of blogging activity that allows fast dissemination of information and exchange of ideas, opinions, and artefacts among diverse audiences. Twitter allows users to interact with microblogs in a variety of ways: sharing it on their profile (retweet), clicking the reaction icon, mentioning someone (by tagging their username), or directly replying to a tweet. Massive amounts of microblogging and interaction data can offer rich insights towards

different issues such as assessing political polarization in the general populace [Conover at al 2021], understanding people's opinions about COVID-19 vaccines [Lyu, Han & Luli 2021], identifying consumers' perceptions of products [Sass et al 2020], and so forth.

In recent years, Twitter has been actively used by scientific communities to exchange scientific information and to promote scholarly discussions [Darling et al 2013]. In particular, Twitter provides a unique opportunity for researchers from various fields to communicate and network with other scientists, disseminate their research findings to the public and media, and engage with a wider audience [Collin, Shifman & Rocks 2016]. Researchers' use of Twitter creates unique opportunities for the public to access cutting-edge research, interact with leading researchers, and learn from them [Martischang et al 2021]. Besides its daily use by scientists, Twitter has been a very powerful tool to enhance communication and interaction in academic conferences beyond the physical spaces of conference rooms [Bombaci et al 2016]. Research shows that many conference participants use Twitter to share ongoing activities and to make quick reflections, and that most remote participants take advantage of conference hashtags in Twitter to follow the event updates and connect with other participants [Shiffman 2012; Collin et al 2016].

As Twitter is increasingly becoming a platform for academic microblogging, more data about the research topics, interests, discussions, and other scholarly interactions are automatically recorded every second. These rich data can provide valuable insights toward understanding the trends and dynamics within the scientific communities [Grover et al 2018]. As a young and fast-growing field, learning analytics can benefit a great deal from such insights. Since its formation in 2011 during the First International Learning Analytics (LAK'11) Conference in Banff, Canada, the field has attracted enormous interest from researchers, practitioners, and policymakers in the last 10 years. As a maturing field of research, the prominent topics in learning analytics have been shifting over time as the researchers continue to explore the ways that learning analytics can create a bigger and sustainable impact in education [Joksimovic et al 2019]. Understanding how the field has been evolving in the last years can help identify the changing research trends, reveal how the community is growing, evaluate the current progress, determine the critical gaps, and plan for the future of the field.

The goal of the present study is to reflect on the ongoing development of the learning analytics field by analyzing the Twitter microblogging data that have been collected between 2010 and 2021. In a past reflection study conducted by Chen and his colleagues [Chen et al 2015], twitter data pertaining to only LAK conferences between 2011-2014 were used. The present research studies a rather comprehensive microblogging dataset to provide an updated reflection on the field.

2 RESEARCH BASED ON TWITTER ANALYTICS

There is a growing number of Twitter-based research. According to [Karami et al 2020] systematic review, Twitter-based research covered a range of topics from application (e.g., politics, health, education) to methodology (e.g., sentiment, big data, topic modeling) and technology (e.g., web technology, digital communication). Studies leveraged Twitter data as a data source to address research questions at both national and international levels. For example, analyzing public conversations about climate change [Veltri & Atanasova 2017; Vu et al 2020], and trends in elections [Buccoliero et al., 2020]. Chen and colleagues [Chen et al 2015] gained insights into the learning analytics community by analyzing tweets that were archived using the official Learning Analytics and Knowledge (LAK) conference hashtags from 2011 to 2014. Their analysis revealed that the tweets can be generally clustered as information related to the conference, personal experiences and comments, and specific research topics. Results of the topic modeling showed that the research topics were becoming more diverse over the years, reflecting the growing field of learning analytics. In

addition, there was an increasing trend of student-centered topics such as assessment, success, engagement, and learning. Besides identifying the topics of interest, the tweets also provided insights into the reach and connections of the LAK community on Twitter.

Schnitzler et al. [Schnitzler et al 2016] argued that Twitter as a social media platform holds a huge potential for researchers to drive research impact. Particularly with the development of new knowledge, approaches, and policies, many research fields are constantly evolving, and hence, researchers can make use of Twitter as a professional activity to increase the visibility and the reach of their work [Cheplygina et al 2020]. Several aspects of Twitter data can be leveraged for research insights [Kim et al 2013; Sinnenberg et al 2017; Williams et al 2013]. One of the key aspects is the digital profile of Twitter users (e.g., the country where the account is registered, number of followers, who the user follows). Costas et al [Costas et al 2017] connected Twitter accounts to names of authors in the Web of Science database to examine the population of scholars from respective fields. The results showed that researchers from the field of Social Sciences and Humanities had the strongest presence on Twitter, and the researchers were also generally younger than those not on Twitter. Mining such information could provide a better understanding of the characteristics underlying a population or subgroup's interest in a topic [Zhou & Na 2019]. The user data also provides research opportunities for exploring connections between users. This can be identified by the retweets and the discussion threads. Using network analysis, the relationship and interactions between Twitter users regarding a specific topic can help identify user influence [Asadi & Agah 2017].

Another aspect is the use of hashtags on Twitter. Hashtags allow users to locate other users based on their interest in similar topics. Therefore, hashtags provide opportunities for interactions, curating resources, and sharing information in an organized manner [Veletsianos 2017]. Kimmons et al [Kimmons et al 2021] examined hashtags that were co-occurring in tweets including the hashtag #EdTech. Their results showed that COVID-19 has influenced the discussions in the field of educational technology with discussions mainly focusing on remote, online, distance, and blended learning. There was also a shift from the emerging use of "remote learning" to "elearning" and "online learning" over time. Furthermore, while the people involved in educational technology on Twitter are from rather broad disciplines, the trends are largely shaped by a small group of highly active Twitter users.

Apart from using hashtags to identify trends, another predominant approach in Twitter-based research has been mining and analyzing the content of tweets [Sinnenberg et al 2017]. Among the different methodologies, topic modeling has been widely deployed to identify underlying concepts discussed in tweets. The utility of this approach is demonstrated by several studies conducted to examine the topics discussed on Twitter about the COVID-19 pandemic (e.g., [Boon-Itt & Skunk 2020; Wicke & Bolognesi 2020]. Prevalent topics that were discussed on Twitter provide insights into the public's perception and awareness of particular domains or constructs of interest. In this paper, exploring the public's discussion surrounding learning analytics could offer useful information as well as alternative perspectives to advance the field

The current study extends previous work [Chen et al 2015] by mining Twitter data related to "learning analytics" from 2010 to 2021. To our knowledge, very few studies have analyzed such a large Twitter dataset specifically on learning analytics. Through this approach, the current study aims to gain a broader understanding of how learning analytics is being discussed in the public space and how it has evolved over the years. To achieve this aim, our analyses will attempt to answer the three main research questions below:

- How have the tweets on learning analytics evolved over the last twelve years based on original tweets and retweets?

- What are the main topics of interest related to learning analytics characterized by the occurrences of Twitter hashtags and the topical words used in Tweets? Are there any shifts in the topics of interest in the last twelve years?
- Where do the most contributors of the Twitter learning analytics discussions originate from?

3 METHODOLOGY

As mentioned earlier, the goal of this study is to perform analysis on a large dataset of Twitter posts (i.e, tweets) with a particular focus on the field of learning analytics from 2010 till 2021. To this end, several steps have been taken sequentially for data cleansing and preparation. After that, three different analysis methods were performed to draw a bigger picture of the field and bring in new insights into the learning analytics community. Figure 1 depicts the universal research workflow. First, we collected the tweets, then we cleaned the data from redundant and unneeded attributes, and after that, the data was prepared for analysis. In this research work, the datasets are processed for three analysis techniques: geocoding, descriptive analysis, and topic modeling. The results from these different analysis steps are discussed in the next section.

In the context of this paper, the term 'tweet' links to a message or post from a Twitter user account, which does not exceed 140-character limit, however, the size of tweets has been extended to 280 characters recently.

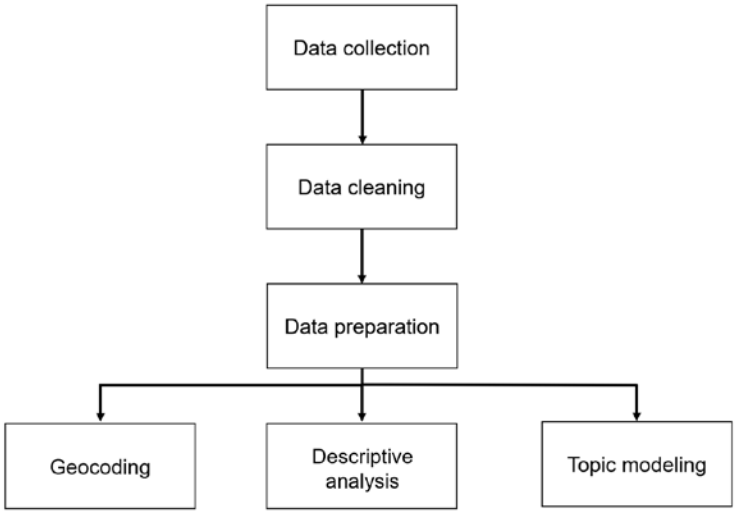

Figure 1: Universal research workflow

3.1 Data collection

The data collection process was performed using the R software. We used twitteR package [Gentry et al 2016] to access the Twitter API and data using private tokens and keys. Most functionalities of the API are supported through the twitteR package commands and the search was optimized to pull out the required raw information (e.g., text, likes, retweets, tweets, users, etc.) to carry out the analyses. The dataset was aggregated by searching for two keywords, "LearningAnalytics" and a bigram form of "Learning Analytics". By searching for these two keywords in a time frame between January 1, 2010, to May 26, 2021, we were able to retrieve 1.616×10^6 tweets posted by 359,216 unique users.

Since the Twitter API has a quota of 900 tweets per 15 mins, the automated process of retrieving the whole dataset of tweets took around 19 days to be stored in our local machines.

3.2 Data cleaning and preparation

As reported by [Chen et al 2015], substantial work is needed to process and clean the tweets before applying a Twitter universal analysis. We used the Natural Language ToolKit (NLTK) library to extract stop words, such as "and", "or", "has", "is", "we", and removed them from the large corpus. We also used Python regex to remove all emojis, which is commonly found in social media posts. In addition, words that are less than 4 letters were excluded. However, and to preserve common short abbreviations related to learning analytics, we created a whitelist that includes specific short words (see section 3.5 for examples). This whitelist was created manually via scanning the top 100 short words from the dataset. We also removed words with low TF-IDF (term frequency-inverse document frequency) as recommended by [Tajbakhsh & Bagherzadeh 2016]. Moreover, empty sentences and duplicate records were removed. At the same time, we customized data preparation and cleaning that fulfil our needs for the geocoding and the topic modeling. The details are described in the sections below.

3.3 Descriptive analysis

Descriptive analysis about the tweets was performed in terms of common hashtags, number of tweets, source of tweets, number of likes, and the language used by the Twitter community on learning analytics from the early stages of 2010 till 2021.

3.4 Geocoding

The second form of analysis conducted was geocoding. Geocoding, or geospatial analysis, is the process of indexing a description of a location such as text or longitude/latitude pair and linking that to a geographical position on a real-world map. Scientists used geocoding and spatial analysis methods to identify trends, explain patterns, and describe various geographical phenomena [Goldberg, 2008]. Although there are different ways to designate a place, such as a selection of predefined places, Twitter relies on locationally descriptive language to denote users' location. For example, some users may state the United States, the others may document Texas as a place of living. Such vernacular self-input texts are easily understood by humans, but they are improper for machines. That is, doing a geospatial mapping of a free text entry by the users requires valid digital data for the computerized environments to understand. Thus, in this paper, a form of technical solutions was followed to convert text descriptors into valid geospatial data.

In the context of this work, we looked at geo-tagged users instead of geo-tagged tweets given that the particular dataset collected from Twitter includes more geospatial data in the user profiles than tweets. In order for us to geospatially display a heatmap of users on the world map, we needed to extract the names of cities and countries from user-input profiles (N= 359,216) and find their geographic location. We used both document parsing for clear semantics and Natural Language Processing techniques to extract city and country names. The latter is used for more complicated wordings; for that purpose, we used spaCy[1]. The pipeline of spaCy includes tokenization of text to parsing till refining a document. To further obtain geographical coordinates by city as well as country names, we decided to deploy a local geocoding service called Nominatinm[2] to find locations on Earth by name and address (i.e., geocoding). Finally, to speed up and simplify the process of geocoding using Nominatinm, we used Nominatim docker[3].

[1] Open-source software library for advanced natural language processing (last accessed August 2021)
[2] Open-source geocoding with OpenStreetMap data (last accessed September 2021)
[3] https://github.com/mediagis/nominatim-docker (last accessed August 2021)

3.5 Topic modeling

The third form of analysis carried out in this study is topic modeling. It is a technique used in discovering hidden semantic structures in a corpus to provide insights into different and common themes [Blei et al 2003]. In the context of this paper, topic modeling is used to identify sub-topics discussed in the collected tweets. To carry out the topic modeling analysis on our dataset and extract main topics from the large corpus, we used Latent Dirichlet Allocation (LDA) [Beli et al 2003], a popular probabilistic topic modeling technique that is commonly used to extract relevant themes and topics in document collections. Briefly, LDA works by exploring word co-occurrences and modeling documents as mixtures of latent topics, which are then modelled as distributions over all possible words. Because of its efficiency and simplicity, LDA becomes one of the key techniques in analyzing large document corpora in the fields of social sciences, humanities, and even social media and newspaper articles [Jelodar et al 2019]. Since LDA is very much affected by the number of topics, we validated the number of topics using a coherence score for the most suitable number of topics [Stevens et al 2012]. Topic coherence calculates the consistency of one topic by measuring semantic similarity between words in a topic.

As the first step of our topic modeling analysis, we cleaned and filtered the tweets (see section 3.2). In addition to the filtration procedure of the tweets, we also:

- built our own two dictionaries for extra inclusion and exclusion of particular words that might prejudice the results of the topic modeling, namely black (e.g., analytics, learninganalytics, analysis, will) and white (e.g., "R, SAS, SRL, IOT, NLP, IIOT, SQL, LMS, IT") lists.
- Removed non-English tweets
- Cleared tweets from hashtags and retweets tags
- Performed lemma using spaCy lemmatization, a pipeline component for assigning base forms to tokens using rules based on part-of-speech tags

After the preparation process, we imported the whole dataset into Python and ran the LDA topic modeling using the Gensim's package[4] and built the topic model.

3.6 Privacy and data protection consideration

A particularly prominent worrisome among the public is whether social media posts should be treated as public or private data [Deacon et al 2021]. Since Twitter is a social networking service with which users cast their thoughts online and to the public, such data (e.g., tweets, likes, quotes, replies) are therefore considered a "public data" [Deacon et al 2021]. Twitter Application Programming Interface (API) prevents mining direct messages and protected/private accounts and requires additional private tokens. Even though tweets might have privacy implications if coupled with publicly available information [Humphreys, Gill & Krishnamurthy 2010], we stress that such engineering has not been performed and the analyses were carried out only on the "public data".

4 RESULTS

4.1 Descriptive analysis

To get a pivotal understanding of the Twitter dataset, we first conducted a general analysis of the dataset in terms of the number of tweets, tweet sources, hashtags used, likes and retweets, as well as languages used. The number of tweets per year was identified from the whole dataset to provide clarity on the level of activity on Twitter across years. Table 1

[4] https://pypi.org/project/gensim/ (last accessed September 2021)

shows a breakdown of the number of tweets per year. It was observed that the number of tweets has steadily increased from 2010 till 2017. In 2018 and 2019, the number of tweets was at the same level as the year 2017. Nevertheless, in 2020 the total number of tweets increased by almost 50% compared to 2019 during which a slight decrease was observed. While the dataset of this study was fetched as of May 2021, the number of tweets in 2021 seems to be rising when compared to the previous years.

Table 1: Count of tweets

Year	Number of tweets	Year	Number of tweets
2010	1,541	2016	153,837
2011	6,849	2017	224,987
2012	16,283	2018	217,168
2013	21,221	2019	213,220
2014	39,730	2020	388,630
2015	73,098	2021	260,334

Next, we examined the data sources of the tweets. Interestingly, we identified 6894 different sources for the collected tweets. Out of the 1.61 x 10^6 collected tweets, we broke down the top 10 sources of the tweets as shown in Table 2. Around three quarters of the top 10 sources belonged to users who tweeted from their computers and smartphone devices, mostly referring to human users. However, and as shown in Table 2, IFTTT is identified as the fifth top source of learning analytics tweets. IFTTT notes "if then else" statement which is commonly used by social media bots that mimic human users. The rest of the tweet sources belong to other devices, bots, and tweet management applications such as TweetDeck.

Table 2: Breakdown of the top 10 sources of the tweets

Source	Number of tweets
Twitter for Android	190,100
Twitter Web Client	164,610
Twitter for iPhone	154,800
Twitter Web App	118,420
IFTTT	74,849
Buffer	55,972
Hootsuite	37,388
TweetDeck	30,919
Twitter for iPad	22,988
MLTweetBotMK	21,280

To get an insight into the number of likes, hashtags, original tweets, and retweets, we tracked frequencies across the years. In total, there are around 1.22 million likes, 1.54 million hashtags used, 0.5 million organic tweets and 1.17 million retweets. Figure 2a shows an increasing number of likes and hashtags during the period of 2011-2015. In the time span of five years between 2015-2020, there is a steep increase in both the likes and hashtags count. Overall, the counts and percentages of the likes and hashtags have increased.

In Figure 2b, we show the number of the original tweets together with the retweets. The descriptive analysis shows that the number of retweets spiked after 2016. Surprisingly, the organic tweets count started to decrease in 2018 up till 2020. Overall, there is an increasing engagement of the learning analytics community on Twitter. While the community

engagement in Twitter was slowly increasing during the first 4-5 years, the peak was achieved in 2020, in which the COVID-19 pandemic hit the world.

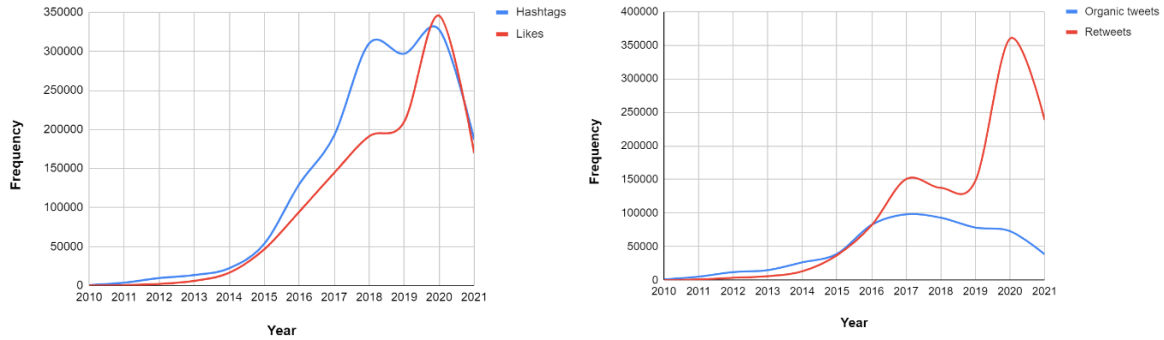

Figure 2a (left): Distribution of the number of Twitter hashtags and likes. Figure 2b (right): Distribution of Twitter organic tweets and retweets

In addition, we looked at the hashtags used in the dataset. In this analysis, #learninganalytics and #analytics were removed as they took dominance in all years. Table 3 shows the top 10 hashtags that were used per year. A few prominent trends could be observed from the top ten hashtags per year shown in Table 3. It is observed that #learning and #edtech were dominant between 2010 and 2016 and became weaker in the later years. From 2012 and 2013 onwards, #machinelearning, #bigdata, and #datascience became increasingly popular. From 2016 onwards, #ai referring to artificial intelligence and #deeplearning gained greater popularity. From 2017 onwards, #iot referring to the internet of things appeared as a dominant hashtag.

Table 3: Top 10 hashtags per year

Year	Hashtag*
2010	#elearning, lak11, learning, jobs, measure, google, edtech, sas, tech, education
2011	#lak11, elearning, lrnchat, learning, edtech, owd11, highered, sa_la, eli2011, kmiou
2012	#bigdata, lak12, edtech, elearning, learning, bi, education, eli2012, highered, socialmedia
2013	#bigdata, learning, edtech, lak13, education, elearning, data,, machinelearning, lasi13
2014	#bigdata, machinelearning, learning, edtech, datascience, elearning, data, azure, laceproject, dalmooc
2015	#bigdata, machinelearning, datascience, learning, deeplearning, python, data, edtech, elearning, azure
2016	#machinelearning, bigdata, datascience, ai, iot, deeplearning, learning, data, ml, edtech
2017	#machinelearning, bigdata, ai, datascience, ml, deeplearning, iot, fintech, tech, artificialintelligence
2018	#machinelearning, ai, bigdata, datascience, deeplearning, ml, iot, python, artificialintelligence, algorithms
2019	#machinelearning, ai, bigdata, datascience, iot, iiot, deeplearning, python, ml, learning
2020	#datascience, bigdata, machinelearning, ai, iot, python, iiot, rstats, pytorch, tensorflow
2021	#datascience, bigdata, machinelearning, ai, python, iot, iiot, rstats, pytorch, 100daysofcode

* Hashtags of #learninganalytics and #analytics are removed

At the final stage of the descriptive analysis and given that the collected data is multilingual, we examined the languages used in the tweet dataset. The most used language in the tweets was the English language as expected (94.99%). We looked at the small proportion (5.01%) of the non-English tweets as shown in Figure 3 and we found that the top used languages after English are Spanish (0.8%), French (0.5%), Dutch (0.37%), and German (0.31%). A very small proportion of the other used languages in the collected tweets belong to Arabic, Turkish, Catalan, Thai, Indian, Chinese,

Portuguese, Korean, Norwegian, Swedish, Danish, Italian, Finnish, Tagalog, and Japanese. These languages count for 26,387 tweets.

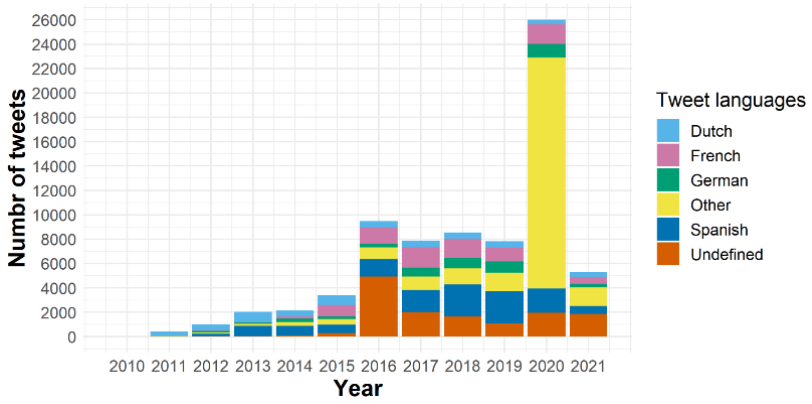

Figure 3: Retrieved tweets count and languages excluding English, n= 7.4×10^4 tweets. Best viewed in color.

4.2 Geocoding analysis

Using geocoding (i.e., geospatial) analysis, we were able to identify 210,420 user origins out of the 359,216 users who tweeted on learning analytics. Because the location is a self-reported option in Twitter, the NLP techniques (see section 3.4) were able to recognize around 60% of the total number of the users. Figure 4a illustrates a heat map of the geotagged users. The results show that some parts of the world have a higher number of users who tweeted on learning analytics. The top 10 countries with the highest number of geotagged users are: United States of America, United Kingdom, India, Canada, Australia, Nigeria, France, Spain, Germany, and the Netherlands. For those who reported their cities in their profiles, the analysis has shown that London is the most reported city of origin of the users, followed by New York, Paris, Bengaluru, and Toronto.

To obtain a more detailed view, we synthesized the data per continent for the utmost tweet users (see Figure 4b). For North America, the east coast of the United States acquired the majority of the users. For example, the cities of New York, Austin, and Boston were among the top 10 US cities. Los Angeles and Seattle were the cities with the highest number of geotagged users on the west coast. For Europe, the highest number of users who tweet on learning analytics origins are from western Europe. The biggest cities like London, Paris, Amsterdam, Barcelona, and Dublin were placed on the top of the cities of origin of the users.

With respect to Africa and Australasia, cities of Bengaluru, Lagos, Mumbai, Sydney, and Melbourne are among the highest reported cities. Other areas of the world like South America have a lower ratio of users. For a full reference of the geocoding analysis, two interactive maps that illustrate the distribution of users by country and a heat map of the Twitter users are available online on a live Jupyter notebook[5].

[5] https://nbviewer.jupyter.org/gist/slate-dev/46a6d784bd3b9566b4087413881b1f28

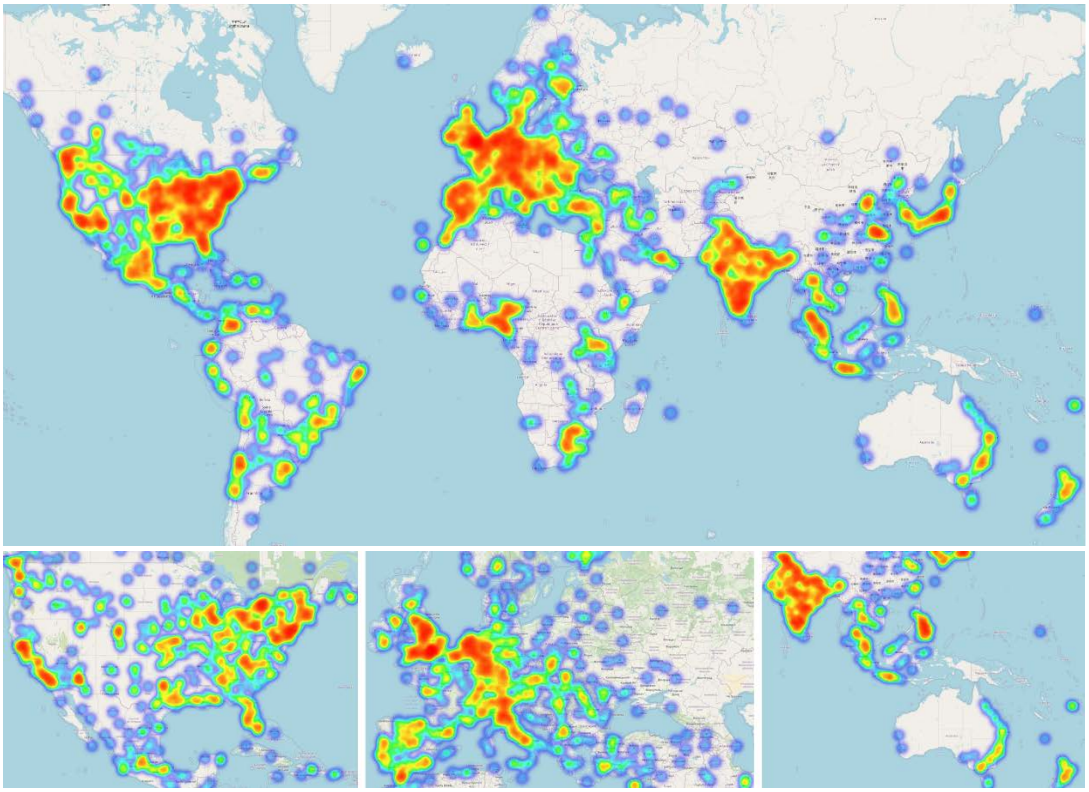

Figure 4: Learning analytics heatmap of Twitter geotagged users. (a) Top, Geotagged users global heatmap. (b) Bottom from left to right, geotagged users heatmap of USA and Canada, Europe, and eastern Australasia

4.3 Topic Modeling

4.3.1 Coherence and the number of topics

To seek a suitable LDA topic number and investigate topics from the Twitter dataset, we conducted several iterations. In each iteration, a coherence score is obtained and later on compared to identify the optimal number of topics [Zhao et al 2015] (see Figure 5).

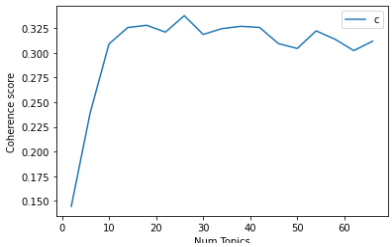

Figure 5: Coherence score for the topic numbers

Using the Gensim's model for measuring the coherence values, the algorithms train more than 68 topics to identify coherence among word co-occurrences for every topic. At the end of the process of training the models, we recognized

that 26 topics is the optimal number for the topics. As seen in Figure 5, the coherence score increases steeply at a value of 0.30 for ten topics. At a number of 26 topics, the coherence value scores the highest at 0.34. In addition, 20-30 topics is a reasonable number of topics for a deep-dive analysis [Dahal et al 2019].

4.3.2 Themes and topics

Based on the similarity among the keywords of the 26 topics, seven different themes were identified. The themes along with their topics and keywords are displayed in Table 4. The themes were labeled by looking at the top 50 words per topic listed according to their occurrences. Three of the authors individually categorized the topics into themes and came together to agree on the final categories used to group the topics, and thereafter, gave an appropriate label to each category. Our results indicate that the majority of the topics include related keywords to "machine learning", "data", "ai", and "science".

The themes that we labeled based on the appearing keywords are: machine learning and machine learning methods including 7 topics, artificial intelligence including 5 topics, information dissemination having 4 topics, education and application evenly having 3 topics per each, and finally social media and privacy and security including 2 topics per each.

Table 4: Topics classification and keywords

Theme, topics	Topic	Keywords*
Theme 1: Machine learning and machine learning methods	T1	data, machine, science, datum, predictive, python, mining, r, kdnugget, platform
	T11	google, machine, cloud, datum, measure, understand, data, training, metric, tool
	T12	machine, text, post, sentiment, library, learner, presentation, help, datum, challenge
	T13	machine, predictive, datum, iot, mobile, ai, model, algorithm, smart, technology
	T14	machine, trend, data, prediction, development, deep, datum, ai, science, research
	T16	datum, machine, data, mining, hadoop, educational, text, science, information, predictive
	T25	machine, deep, algorithm, intelligence, artificial, search, data, design, computer, vision
Theme 2: Artificial Intelligence	T3	machine, ai, capability, predictive, data, technology, intelligence, advanced, artificial, build
	T5	machine, ai, service, datum, human, system, platform, solution, technology, deep
	T6	machine, intelligence, azure, artificial, business, predictive, open, data, source, transform
	T17	deep, machine, discuss, framework, report, datum, predictive, graph, tensorflow, personalized
	T18	machine, model, problem, ai, deep, infographic, share, work, ml, recognition
Theme 3: Information dissemination	T2	start, free, book, machine, online, time, join, director, announce, sign
	T10	data, machine, great, practice, answer, path, thank, presentation, meeting, resource
	T23	webinar, look, machine, join, register, interested, event, conference, live, morning
	T24	education, student, research, university, support, conference, keynote, workshop, talk, summit
Theme 4: Education	T7	time, spend, work, career, enjoy, machine, college, student, goal, school
	T9	student, improve, help, experience, design, engagement, track, online, teacher, elearning
	T20	machine, mooc, think, game, teaching, work, classroom, present, predictive, talk
Theme 5: Application	T0	insight, application, deep, spark, language, engineer, enterprise, processing, natural
	T21	machine, datum, data, product, ai, business, cloud, oracle, solution, chain
	T22	machine, datum, digital, technology, automate, impact, revolution, case, poll, operation

Theme 6: Privacy and Security	T4	machine, risk, threat, data, reduce, cybersecurity, help, jisc, code, practice
	T19	machine, network, deep, datum, difference, detection, security, fraud, case, attack
Theme 7: Social Media	T8	future, machine, blog, video, skill, management, post, late, free, read
	T15	social, medium, marketing, firm, digital, facebook, sentiment, follower, reader, analysis

* Keywords are sorted according to their relevance in the context of each topic

At the last stage of our analysis, we examined the changes in topic frequency over the past 11 years. Figure 6 shows a line graph which depicts the relevance metric (i.e., topic weight) of each theme in the time period between 2010-2021. According to Figure 6, the Security, Application, and Artificial Intelligence themes, which were less popular during the initial years of the field, show a steady increase, whereas the Education and Information Dissemination themes decline continuously after their increasing popularity in the first two years of the field. The Machine Learning theme shows a stable trend throughout the years. The Social Media theme was quite popular in 2010, but its popularity continuously decreased until 2015. After 2015, this theme was stable until 2021, where it showed a significant jump.

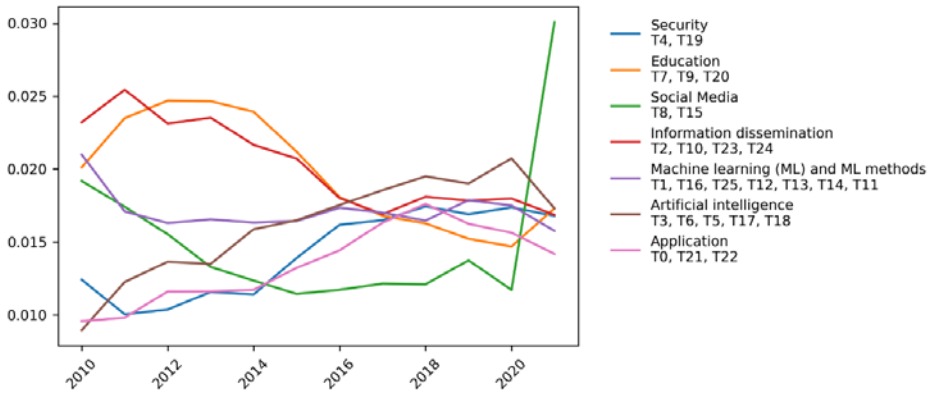

Figure 6: Changing in learning analytics tweet theme distribution over the time span 2010-2021. Best viewed in color.

5 DISCUSSION

The current study leveraged public discussions on Twitter surrounding learning analytics to examine the development of the field over the last ten years. Unstructured public discourse offers wider perspectives beyond discussions within the scientific community, thus, providing an overarching understanding of what is being discussed and who or where are the sources of influence in the Twittersphere. In a recent paper by Selwyn [Selwyn 2020], Twitter discussions on learning analytics were highlighted as examples of concerns about how data is being used as surveillance instead of support. It appears that Twitter has provided a much-needed space for dynamic discussions and participation across groups (e.g., policy makers, teachers, students, researchers). The current analysis points to three noteworthy findings which respectively answer our three research questions.

Firstly, learning analytics is a maturing field where discussions have extended beyond the selected few. The increasing number of hashtags and tweets indicate a steady growth in interest in learning analytics. As an emerging field in 2011, there were more organic tweets by individual users. As the field progressed, there were more retweets suggesting a developing network of members in the learning analytics community with ideas that are spreading out from individuals. In parallel with Peri et al.'s study [Peri et al 2020], where epidemiological methods were used to

measure engagement patterns in Twitter, the connections and influence of highly infective individuals can be used to increase greater social participation. Aligned with Chen et al.'s findings [Chen et al 2015], the topics of interest in the learning analytics community become more diverse as indicated by the increasing number of hashtags being used.

With respect to the second research question, the results provided some insights into the changes in trending topics in learning analytics. The changes in the popularity of hashtags indicate the interest of the field in the advancement of new technologies. Given that a big part of learning analytics is on making sense of large sets of data, certain programming languages and analysis tools and expertise are needed for the analytics. This is demonstrated by the popularity of the hashtags #python and #rstats in more recent years and matches the calls for required expertise to establish learning analytics as services [Arnold et al 2014].

Besides the shift in popularity of certain hashtags, the dominance of some hashtags in some particular years appeared to be related to conferences, namely the Learning Analytics and Knowledge (LAK) as well as Learning Analytics Summer Institute (LASI). Those appeared with dominance at the early stages of the field of learning analytics, such hashtags were (#lak11, #lak12, #lak13, #eli2011, #eli2012, #owd11, #lasi13). The dominance of conference-related hashtags might indicate that discussions and sharing of ideas about learning analytics are most active during conferences [Chen et al 2015]. Another type of hashtags that are dominant in a specific year appeared to be related to various initiatives in the community. For example, #laceproject referring to learning analytics community exchange, #dalmooc referring to data, analytics, and learning MOOC, and #lrnchat was used to bring people who are interested in discussing topics related to learning together.

Moreover, the topic modeling resulted in 26 topics, which were then organized into seven themes. Among the seven themes, two of which are related to Machine Learning and AI, with a total of 12 topics, 5 topics to the prior and 7 to the latter. This result indicates the strong interest of the community in exploring advanced and opaque technologies and techniques. These themes were identified to have a changing popularity since the beginning of the field, yet with an increasing interest. This comes as no surprise given that a great focus of the field of learning analytics has been shifted towards predictive modeling and the uses of artificial intelligence to support students at risk and personalize interventions [Prinsloo & Slade 2017].

Security and Application themes have an increasing trend since the initial years of the field emergence. This result may show that in parallel to the increasing applications of artificial intelligence technologies in education, there has been a growing discussion on the privacy and security issues. The appearance of the security theme, which involves privacy and ethics keywords, suggests increasing interest in those aspects of the learning analytics field. One might argue that considering the security issues and the data protection concerns on the social media sphere is part of the learning analytics discourse [Greller & Drachsler 2012] and the discussion continues outside the academic venues.

As over the years learning analytics matured, Education and Information dissemination themes are noted to have a decreasing trend, which could indicate that the community focuses less on the educational part of the field and the hype of the new emergent field in the period of 2011-2014 dissolves with other rising topics such as machine learning and artificial intelligence. These findings shed considerable light on how the community has been increasingly discussing the advanced technologies used in the field at the same time as they have been talking less about the use and effects of them in education. Perhaps such a trend should warrant caution and raise considerable attention since the field of learning analytics has a data-driven nature and establishing a direct impact on education presents a central challenge [Wiley et al 2020].

Further inspection of the topics identified an interesting finding through the presence of the Social Media topics in 2011 and the abrupt jump in the recent two years. Strongly related to the start of the COVID-19 pandemic, this may

suggest that there are increasing discussions on related work of learning analytics in the online sphere. This aligns with the recent findings on the increased usage of social media and information during the pandemic [Tsao et al 2021]. As an implication, we foresee that there is more interest in leveraging such data in the field of learning analytics through the appearance of words like "sentiment" and "analysis" in topic 8 and topic 15.

Finally, and to answer our third research question "Where do the most contributors of the Twitter learning analytics discussion originate from?", the geocoding analysis revealed interesting findings. Even though the field of learning analytics has an international reach and is feasibly seeking to shape research, opinions and practices across the globe, we noticed a distinctive Global North sharing a wider range of tweeting users than users from the Global South countries, with the exception of India. Perhaps an explanation for such finding is because Twitter is found most popular in the USA and Europe [Java et al 2007]. Other explanations might relate to the socioeconomic gradient, technological advancement, internet access, and the higher educational level in the Global North.

Also, while most of the tweet users were found to be from the Global North, we discovered that some regions of the Global North are not as active as the others. For example, Eastern Europe and the Scandinavian countries were not found as prevalent as those from particular countries from Western Europe, namely the UK, Germany, and France. One explanation is that the small population in some of these countries affect the overall quota of the users in the dataset.

In addition, the exception of India and the rising number of Twitter users from other parts of the world such as Nigeria suggests that the field has become more popular and diverse. Perhaps Twitter users from those countries are more likely to post and use social media features such as likes and retweets in order to join important discussions that are relevant to their interests and education. Nevertheless, the explanation of the existence of Twitter bots is rational in our dataset, given that some of the data sources were found commonly used by bots which could account for a large chunk of the traffic produced.

6 LIMITATIONS

Carrying out a study based on social network services has limitations. With specific focus on Twitter, there are limitations that could have affected the present study and the conclusions of this work. First, even though the crawled tweet dataset is large, it is worthy to note that some tweets could have been lost and not mined by the Twitter API which by then could affect the general comprehensive of the archive. Second, while the search term used in the study has particularly crawl "learning analytics" as a singular and bigram semantic, perhaps there are tweets that are irrelevant to the context of learning analytics which influences our spatial, descriptive, and topic modeling analyses. Third, our manual investigation suspected some Twitter users and flagged them as bots. Discovering social media bots is challenging since bots are aligning with human activity on Twitter and becomes more and more difficult to differentiate from genuine user accounts [Ferrara 2020]. In addition, while we tried to show the geographical locations of the users who tweet on learning analytics, it was remarkable that some countries were underrepresented such as China and Russia on the Twitter sphere (for possible causes, see [Benney 2011]. Fourth, the topic modeling and the LDA have limitations in terms of lack of nuances for qualitative thematic analysis. For instance, we identified the common themes from the topics subjectively which might have created bias in the results.

As a future direction for this study, we plan to improve the tweet filtration process by identifying and excluding Twitter bots as well as link the tweets to academic publications of learning analytics. Furthermore, we predict that such a procedure might present finer conclusions. We also plan to normalize the total number of tweets per country as a mean to provide a better validation of the country representation in the geospatial analysis.

7 CONCLUSION

The current study built on previous work to examine the evolution of the field of learning analytics by mining Twitter data with learning analytics as keywords. Given that learning analytics is not only a multidisciplinary field, but also a field that needs to factor in perspectives of students, teachers, and policy makers, this approach allowed us to examine a wider public view on learning analytics beyond Tweets tagged to conferences or specific topics. In recent years, the advancement of new technologies and programming languages appears to have shaped the discussions on Twitter, suggesting the important role learning analytics play in understanding how data analytics can inform and advance education enhanced by emerging technologies.

We also conclude that the field of learning analytics might strive for an additional global perspective on the social network services realm. That is, the community should seek more ways to keep connected and increase diverse opinions. We may then agree with what [Shum & Luckin 2019] who proposed that the future might bring unique practices and clashing politics from certain parts of the world which are barely represented in the community of learning analytics.

The last years have been an exciting period of growth and experimentation for the field of learning analytics. We cannot help but wonder what the next ten years would be like and what are the active steps that we, as a learning analytics community, can take to engage a wider population and further our discussions beyond academic conferences.

REFERENCES

Ahlgren, M. (2020). 50+ Twitter Statistics & Facts For 2020. Accessed: Mar. 22, 2020. [Online]. Available: https://www.websitehostingrating.com/twitter-statistics/

Arnold, K. E., Lynch, G., Huston, D., Wong, L., Jorn, L., & Olsen, C. W. (2014). Building institutional capacities and competencies for systemic learning analytics initiatives. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 257–260. https://doi.org/10.1145/2567574.2567593

Asadi, M., & Agah, A. (2017). Characterizing user influence within Twitter. *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 122–132.

Benney, J. (2011). Twitter and Legal Activism in China. *Communication, Politics & Culture*, *44*(1), 5–20. https://doi.org/10.3316/informit.127165324261101

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bombaci, S. P., Farr, C. M., Gallo, H. T., Mangan, A. M., Stinson, L. T., Kaushik, M., & Pejchar, L. (2016). Using Twitter to communicate conservation science from a professional conference. *Conservation Biology*, *30*(1), 216–225.

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, *6*(4), e21978.

Buccoliero, L., Bellio, E., Crestini, G., & Arkoudas, A. (2020). Twitter and politics: Evidence from the US presidential elections 2016. *Journal of Marketing Communications*, *26*(1), 88–114.

Chen, B., Chen, X., & Xing, W. (2015). "Twitter Archeology" of learning analytics and knowledge conferences. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 340–349.

Cheplygina, V., Hermans, F., Albers, C., Bielczyk, N., & Smeets, I. (2020). *Ten simple rules for getting started on Twitter as a scientist*. Public Library of Science San Francisco, CA USA.

Collins, K., Shiffman, D., & Rock, J. (2016). How are scientists using social media in the workplace? *PloS One*, *11*(10), e0162680.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Fifth International AAAI Conference on Weblogs and Social Media*.

Costas, R., van Honk, J., & Franssen, T. (2017). Scholars on Twitter: Who and how many are they? *ArXiv Preprint ArXiv:1712.05667*.

Dahal, B., Kumar, S. A. P., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, *9*(1), 24. https://doi.org/10.1007/s13278-019-0568-8

Darling, E. S., Shiffman, D., Côté, I. M., & Drew, J. A. (2013). The role of Twitter in the life cycle of a scientific publication. *ArXiv Preprint ArXiv:1305.0435*.

Deacon, D., Pickering, M., Golding, P., & Murdock, G. (2021). *Researching communications: A practical guide to methods in media and cultural analysis*. Bloomsbury Publishing USA.

Ferrara, E. (2020). # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*.

Gentry, J., Gentry, M. J., RSQLite, S., & Artistic, Rm. L. (2016). Package 'twitteR.' *R Package Version*, *1*(9).

Goldberg, D. W. (2009). A Geocoding Best Practices Guide, The North American Association of Central Cancer Registries. *Http://Www. Naaccr. Org/Filesystem/Pdf/Geocoding_Best_Practices. Pdf*.

Greller, W., & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Journal of Educational Technology & Society*, *15*(3), 42–57.

Grover, P., Kar, A. K., & Davies, G. (2018). "Technology enabled Health"–Insights from twitter analytics with a socio-technical perspective. *International Journal of Information Management*, *43*, 85–97.

Humphreys, L., Gill, P., & Krishnamurthy, E. (2010). *PRIVACY ON TWITTER 1 How much is too much? Privacy issues on Twitter.*

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 56–65. https://doi.org/10.1145/1348549.1348556

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211.

Joksimović, S., Kovanović, V., & Dawson, S. (2019). The journey of learning analytics. *HERDSA Review of Higher Education*, *6*, 27–63.

Karami, A., Lundy, M., Webb, F. & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. *IEEE Access*, *8*.

Kim, A. E., Hansen, H. M., Murphy, J., Richards, A. K., Duke, J., & Allen, J. A. (2013). Methodological considerations in analyzing Twitter data. *Journal of the National Cancer Institute Monographs*, *2013*(47), 140–146.

Kimmons, R., Rosenberg, J., & Allman, B. (2021). Trends in educational technology: What Facebook, Twitter, and Scopus can tell us about current research and practice. *TechTrends*, 1–12.

Lyu, J. C., Le Han, E., & Luli, G. K. (2021). COVID-19 vaccine–related discussion on Twitter: Topic modeling and sentiment analysis. *Journal of Medical Internet Research*, *23*(6), e24435.

Martischang, R., Tartari, E., Kilpatrick, C., Mackenzie, G., Carter, V., Castro-Sánchez, E., Márquez-Villarreal, H., Otter, J. A., Perencevich, E., & Silber, D. (2021). Enhancing engagement beyond the conference walls: Analysis of Twitter use at# ICPIC2019 infection prevention and control conference. *Antimicrobial Resistance & Infection Control*, *10*(1), 1–10.

Peri, S. S. S., Chen, B., Dougall, A. L., & Siemens, G. (2020). Towards understanding the lifespan and spread of ideas: Epidemiological modeling of participation on Twitter. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 197–202. https://doi.org/10.1145/3375462.3375515

Prinsloo, P., & Slade, S. (2017). An elephant in the learning analytics room: The obligation to act. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 46–55. https://doi.org/10.1145/3027385.3027406

Sass, C. A. B., Pimentel, T. C., Aleixo, M. G. B., Dantas, T. M., Cyrino Oliveira, F. L., de Freitas, M. Q., da Cruz, A. G., & Esmerino, E. A. (2020). Exploring social media data to understand consumers' perception of eggs: A multilingual study using Twitter. *Journal of Sensory Studies*, *35*(6), e12607.

Schnitzler, K., Davies, N., Ross, F., & Harris, R. (2016). Using Twitter™ to drive research impact: A discussion of strategies, opportunities and challenges. *International Journal of Nursing Studies*, *59*, 15–26.

Selwyn, N. (2020). Re-imagining 'Learning Analytics' … a case for starting again? *The Internet and Higher Education*, *46*, 100745. https://doi.org/10.1016/j.iheduc.2020.100745

Shiffman, D. S. (2012). Twitter as a tool for conservation education and outreach: What scientific conferences can do to promote live-tweeting. *Journal of Environmental Studies and Sciences*, *2*(3), 257–262.

Shum, S. J. B., & Luckin, R. (2019). Learning analytics and AI: Politics, pedagogy and practices. *British Journal of Educational Technology*, *50*(6), 2785–2793.

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, *107*(1), e1–e8.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.

Tajbakhsh, M. S., & Bagherzadeh, J. (2016). Microblogging hash tag recommendation system based on semantic TF-IDF: Twitter use case. *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 252–257.

Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: A scoping review. *The Lancet Digital Health*, *3*(3), e175–e194. https://doi.org/10.1016/S2589-7500(20)30315-0

Veletsianos, G. (2017). Three cases of hashtags used as learning and professional development environments. *TechTrends*, *61*(3), 284–292.

Veltri, G. A., & Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, *26*(6), 721–737.

Vu, H. T., Do, H. V., Seo, H., & Liu, Y. (2020). Who Leads the Conversation on Climate Change?: A Study of a Global Network of NGOs on Twitter. *Environmental Communication*, *14*(4), 450–464.

Wicke, P., & Bolognesi, M. M. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS One*, *15*(9), e0240010.

Wiley, K. J., Dimitriadis, Y., Bradford, A., & Linn, M. C. (2020). From theory to action: Developing and evaluating learning analytics for learning design. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 569–578. https://doi.org/10.1145/3375462.3375540

Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Zhou, Y., & Na, J. (2019). A comparative analysis of Twitter users who Tweeted on psychology and political science journal articles. *Online Information Review*.