Conspiracy-BERT: A Pre-Trained Language Model for Conspiracy Theory Tweets

Ryan Cooper
College of Computing
Georgia Institute of Technology
ryan.cooper@gatech.edu

Thor Keller
College of Computing
Georgia Institute of Technology
tkeller34@gatech.edu

Christian Boylston
College of Computing
Georgia Institute of Technology
cboylston@gatech.edu

Jonathan Leo
College of Computing
Georgia Institute of Technology
jpleo122@gatech.edu

Abstract

In this paper, we introduce Conspiracy-BERT, a transformer-based model pre-trained on a corpus of over 12 million tweets related to various conspiracy theories such as QAnon, Flat Earth theory, and the COVID-19 "plandemic." Our model shows a relative improvement of over 23% when compared to its base model BERT-Large on the task of conspiracy stance detection in tweets. We believe our language model optimized on popular conspiracy content will be an invaluable tool for downstream NLP tasks related to conspiracies for content moderators and researchers looking to curb the spread of dangerous conspiracies online. Additionally, we contribute a large unlabeled data set of tweets related to 12 different conspiracies and a data set of 5000 conspiracy tweets labeled for stance detection.

1 Introduction

The rise of social media platforms has facilitated the democratization of information where everyday citizens no longer derive what is true exclusively from the authority of major publishers like newspapers or cable news. Now, truth exists as a consensus determined by conversation amongst online peers. These developments have opened the doors for some incredible citizen journalism and a wider representation of different viewpoints, but have also given a voice to those who spread verifiable falsehoods paraded as truth. Stopping the spread of misinformation has become a major concern of online platforms with sites like Facebook and Twitter going so far as to hire independent fact-checkers and create content warning tags for posts containing potential misinformation. Misinformation comes in a lot of forms, but conspiracy theories have proven to be a surprisingly potent and dangerous subcategory. To improve understanding surrounding conspiracies on Twitter we have developed the ConspiracyBERT model [1].

1.1 Motivation

Over the past several years, online conspiracy theories have frequently led to real-life danger. The QAnon conspiracy was labeled as a domestic terrorism threat by the FBI (Zeke Miller and Seitz [2020]), Pizzagate compelled a man to shoot up a pizza parlor (Haag and Salam [2017]), and Sandy Hook false flag theories resulted in the continued harassment of the families of the massacre's victims

[1] https://github.com/DLT-FALL-2020/ConspiracyBERT

(Allen [2020]). Given these events and many similar ones, platforms have been under increasing pressure to curb the spread of the conspiracies that have flourished on their platforms. Many have gone to great lengths to combat this content, with Youtube reducing recommendations for conspiracy videos and Twitter even banning thousands of accounts associated with the QAnon conspiracy theory (Conger [2020]). Despite these efforts, new conspiracies continue to form and spread on the platform with recent examples spawning from COVID-19 and the roll-out of 5G cellular service (Ahmed et al. [2020]). This has proven to be a particularly salient issue for decentralized platforms like Twitter where conspiracy theorists can broadcast to the world rather than be relegated to communities of like-minded individuals; posing a real and severe threat to things like public health with misinformation regarding COVID-19. Sites like Reddit have been able to suppress the spread of conspiracies by banning conspiracy theory forums (Robertson [2018]), but moderating the ever-flowing cascade of new conspiracies on decentralized platforms remains an open problem. Currently, few tools are openly available to help these platforms with most current research more focused on modeling the spread of conspiracies post-hoc.

1.2 Related Work

Academic interest in conspiracy theories has long predated the advent of social media platforms. As such, the lion's share of the literature on conspiracies is in regards to the psychology profiles of conspiracy theorists. Some of the biggest insights from these works are that belief in one conspiracy is highly correlated with belief in others suggesting that conspiracy belief may be a monological belief system (Swami et al. [2010]), those who are anxious, socially isolated, and feeling powerless are more likely to be drawn to conspiracies (Goreis and Voracek [2019]), and conspiracies may serve "to uphold the image of the self and the in-group as competent and moral but as sabotaged by powerful and unscrupulous others" for believers (Douglas et al. [2017]).

There has been some work in the computational social sciences to support these claims. One study of the linguistic features of conspiracy theorists on Reddit conspiracy communities found that conspiracy believers were likely to fixate on authority/hierarchical structures using words like "law," "power," and "government" more so than other reddit users (Klein et al. [2018]). Moreover, they were more concerned with language surrounding abuses of power such as "crime," "stealing," and "deception" (Klein et al. [2018]). Other computational work has attempted to model the narratives of conspiracies online using entity and relational extraction from Reddit data and news articles (Tangherlini et al. [2020]). In addition, several studies have studied the spread of COVID-19 misinformation and conspiracies on social media platforms (Alam et al. [2020], Huang and Carley [2020], Ferrara [2020]).

Little work has been done to leverage recent advances in NLP, like transformer models, in the domain of conspiracy theories. There has been success in applying these language models such as BERT (Devlin et al. [2018]) to specific domains like scientific papers (Beltagy et al. [2019]) and biomedical (Lee et al. [2019]) papers for increased performance on relevant downstream NLP tasks. The most similar domain to conspiracies that has been explored is COVID-Twitter-BERT (Müller et al. [2020]), which was trained on a corpus of tweets related COVID-19 and showed increased performance on tasks related to COVID-19 when compared to it's base model BERT-Large.

1.3 Research Goal and Contributions

Given the current research in online conspiracy theories suggesting that proponents of online conspiracy theories have unique vocabularies, motivations, and fixations on certain narratives and figures, we believe that language modeling specific to the conspiracy domain is a useful contribution in better understanding, modeling, and moderating the monological believe system of conspiracy theorists as it presents itself in their use of language. We've seen success in the use of domain-adapted language models such as COVID-Twitter-BERT for downstream tasks related to COVID-19 or vaccines and present our own model adapted to the domain of conspiracy theory tweets.

In order to do this, we constructed a data set of conspiracy theory tweets related to the following 12 conspiracy topics: COVID-19, 9/11 truth, QAnon, Space (Flat Earth/Moon Landing), New World Order, Vaccines, Sandy Hook, Clinton Familiy, Climate (Chemtrails/Global Warming), Jeffrey Epstein, White Nationalism, and Alex Jones. We trained BERT from scratch three times on this data set initializing with the weights of BERT-Base, BERT-Large, and Covid-Twitter-BERT. Our goal is that by training over a variety of different conspiracy across a long period of time our models will be generalized models of the langauge surrounding conspiratorial beliefs on Twitter. We hope our models can be used on a variety of downstream tasks related to conspiracies such as conspiracy detection, entity extraction, and relational extraction. Additionally, we contribute our large unlabeled data set of tweets across the 12 conspiracies along with a subset labeled for whether they promote the conspiracy or not. We hope these data sets can be used in future work for studying conspiracies online.

2 Methods

2.1 Data Collection

In order to create ConspiracyBERT, we needed to construct a conspiracy related data set. Using the MIT Media Cloud Explorer[2], we reviewed the most common article titles and entities related to conspiracies. From these we constructed topical categories for the conspiracies we found such as QAnon or Alex Jones. Once this was complete we decided to focus on the following 12 conspiracy topics: COVID-19, 9/11 truth, QAnon, Space (Flat Earth/Moon Landing), New World Order, Vaccines, Sandy Hook, Clinton Familiy, Climate (Chemtrails/Global Warming), Jeffrey Epstein, White Nationalism, and Alex Jones.

Knowing the conspiracies we were focusing on, we needed to collect Twitter data related each. To do this we decided to construct a list of hashtags related to each conspiracy category and collect all tweets containing those hashtags. We choose this approach because hashtags are easy to identify and are generally strong indicators of topics for tweets (Mehrotra et al. [2013]). To begin, for each conspiracy we determined through manual review a single hashtag for each conspiracy that promoted the conspiracy. This would generally be something like "#QAnon" for the QAnon conspiracy or "#911Truth" for 9/11 related conspiracies. We can see the starting hashtag for each conspiracy in Table 1.

Conspiracy	Starting Hashtag	Co-occuring Hashtags Examples	# of Hashtags	Tweets Collected
Alex Jones	#alexjones	#infowars, #infowarsarmy, #prisonplanet	6	1,431,535
Climate	#chemtrails	#geoengineering, #weatherwarfare, #globalwarminghoax	32	1,464,644
Clinton Family	#clintonbodycount	#clintoncrimefamily, #sethrichcoverup, #clintoncabal	18	427,606
COVID-19	#plandemic	#5gcoronavirus, #fakevirus, #billgatesbioterrorist	42	123,731
Jeffrey Epstein	#epsteindidntkillhimself	#epsteincoverup, #epsteinbodydouble, #epsteinmurder	11	151,417
New World Order	#newworldorder	#nwo, #illuminati, #killuminati	22	3,459,733
Sandy Hook	#sandyhookhoax	#sandyhoax, #noahposner, #crisisactors	6	20,796
Space	#flatearth	#researchflatearth, #globetards, #nasalies	39	661,594
QAnon	#qanon	#wwg1wga, #qarmy, #deepstate	68	2,607,355
White Nationalism	#whitegenocide	#14words, #stopwhitegenocide, #waronwhites	9	262,832
Vaccines	#vaccineskill	#vaccineagenda, #vaxxed, #vaccinesarepoison	27	20,796
9/11 Truth	#911truth	#building7, #911wasaninsidejob, #controlleddemolition	21	186,349

Table 1: Hashtag Snowball Sampling

With the starting hashtags for each conspiracy established, we collected all tweets containing the hashtag in question leaving us with 12 tweet data sets (i.e. one for each conspiracy). In order to further populate our tweet data sets for each conspiracy, we had to find more relevant hashtags for each conspiracy and collect tweets for those. To do this, we followed a snowball sampling method for hashtags. After collecting all tweets containing the starting hashtag (e.g. "#911Truth" or "#QAnon") for a given conspiracy, we quantitatively collect the 100 most commonly co-occurring hashtags in the data set excluding our starting hashtag. For each of these 100 hashtags, we collect 10 tweets randomly that contain them from the Twitter Search API. Then if over 90% of those tweets were determined by the reviewers to be related to the conspiracy in question the hashtag was added to to the list of related hashtags for that specific conspiracy. We then collect all the tweets that contain the new hashtags for a given conspiracy. We repeat this process of finding co-occurring hashtags in the data set (excluding hashtags that have already been collected in the co-occurrence labeling phase) and collecting tweets containing relevant ones until we have no new hashtags to collect, at which point our data collection is complete. When data collection was complete we had a total of 12,294,650 tweets. In Table 1 we can see an example of the starting hashtag, the total number of hashtags collected, the total number of tweets collected and some of the most frequently co-occurring hashtags for each conspiracy.

[2] https://explorer.mediacloud.org/

2.2 Dataset Labeling

In order to validate the performance of our model on relevant moderation tasks, we constructed a data set of tweets that were labeled as promoting conspiracy or not. To construct this binary stance classification data set, we took a stratified sample of 5000 tweets across all of the 12 conspiracies. By stratified, we mean that if say QAnon related tweets make up 20% of all of our tweets collected than 20% of the 5000 tweet sample should be QAnon related tweets. We did this to make sure the model performance was robust across a variety of conspiracies and not just the most prevalent in the data set like QAnon and New World Order. Once the sample was collected we then deleted these tweets from the broader data set along with any duplicates in other conspiracy categories to ensure the model isn't pre-trained on data with which it will be validated. The criteria that we used for labeling was to only label the tweets as conspiracy promoting if they explicitly expressed a positive stance towards the broader conspiracy. If we were unsure whether a post was a joke, we would consult the user's profile page and come to a conclusion through discussion. If we still remained unsure we labeled the data as not promoting conspiracy. Below in Table 2 one can see some tweets that we would label as 1 (conspiracy promoting) or 0 (not conspiracy promoting). Our final data set contained 5000 tweets with roughly 70% promoting conspiracies.

Tweet	Label (0/1)
The #DeepStateCabal now on a crusade to get a rise out of Americans. Things don't add up in Minneapolis. There was a recent Q drop	1
Chemtrails accelerated over America in late 1990's. Cries of global warming fear mongering/ tipping point began 1989.	1
I remember when #AlexJones used to talk about chemtrails and lizard people. Time to go back to your roots.	0
Kim Kardashian actually had twins, but she sacrificed the other one to Satan #illuminati #TheMoreYouKnow	0

Table 2: Example Conspiracy Stance Tweets and Labels

2.3 Data Cleaning

Some conspiracies have content that bleeds together given the interconnected nature of the conspiracy universe. This can result in tweets that reference both conspiracies at the same time. Since our data collection follows a snowballing method these tweets can end up in two separate conspiracy categories, which can add noise to training. To overcome this, all duplicate tweets were removed from the data set and labeled according to the first category they were included in.

Before any type of training, there were some cleaning steps taken before inputting the data in the model. All instances of emojis were removed from the dataset as they would add unneeded noise during training. Urls were also removed from the dataset since the model doesn't have the ability to judge the content of the linked entity and the BERT tokenizer was not designed to handle URLs.

2.4 Model Training and Evaluation

ConspiracyBERT is trained to detect whether a tweet is for or against a conspiracy. To accomplish this, ConspiracyBERT undergoes two rounds of training starting from prebuilt BERT models, this architecture can be seen in Figure 1. Given a pretrained BERT model, the model is further pretrained on a domain specific data set, which in our case is our unlabeled conspiracy data set, consisting of 12 million tweets categorized by conspriacy. Next this model is fine-tuned on our labeled conspiracy stance data set to identify conspiracy stance. Each task has an 80-20 data train test split and uses the recommended parameters, optimizations, and loss functions from the initial BERT paper (Devlin et al. [2018]).

The overall goal of the pretraining step is for the model to hone in on conspiratorial language. Since the initial models used as input for this step have been trained as general language models, our idea was to pretrain these models on a conspiracy tweet corpus. We provide this from our data collection efforts, where a dataset containing 12 million tweets categprized by their conspiracy category is yielded. To capture conspiratorial language, pretraining is formatted as a multiclass single sentence classification task where each conspiracy is it's own class. According to this task, given a tweet the

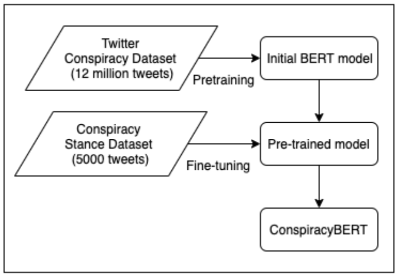
Twitter Conspiracy Dataset (12 million tweets)
Pretraining
Initial BERT model
Conspiracy Stance Dataset (5000 tweets)
Fine-tuning
Pre-trained model
ConspiracyBERT

Figure 1: Model Training Architecture

model must predict the tweet's conspiracy class. Normally, BERT downstream pretraining tasks are unsupervised, but since we had data already labeled, we decided to supervise pretraining according to these conspiracy labels.

Once the model has been pretrained over conspiratorial tweets, the model is further fine-tuned so it captures tweet conspiracy sentiment, which is the end goal for ConspiracyBERT. The fine-tuning task is a binary stance classification task, which stems from the labeled conspiracy stance dataset we created. This dataset contains 5000 tweets from a stratified sampling of the larger Conspiracy Twitter dataset, so that each conspiracy category is proportionally represented similarly. Each tweet is labeled either as promoting or not promoting it's conspiracy category and then removed from the Conspiracy Twitter dataset. Thus the goal of the model during fine-tuning is given a tweet, predict the tweet's conspiracy stance.

In order to see the effect of the base language model on the classification accuracy of the downstream task, we will use three different BERT-based models as the source model for pretraining: BERT-base, BERT-large, and Covid-Twitter-BERT (Müller et al. [2020]). In addition to the evaluating on the dataset we introduce, we also compared each of our models to the datasets used in the Covid-Twitter-BERT. These data sets include: COVID-19 Category (CC), Stanford Sentiment Treebank v2 (SST-2), and Twitter Sentiment SemEval (SE). CC is a binary classification task on COVID-19 tweets classifying if the tweets are news (66.7%) or personal narrative (33.3%). SST-2 is movie reivew binary classification dataset classifying the reviews as positive (55.7%) or negative (44.3%). Finally, SE is Twitter sentiment classification dataset with labels for negative (15.7%), neutral (45.9%) and positive (38.4%) comments.

3 Experiments and Results

Model	F1 Weighted	Training Time (DGX-2)
BERT-Base	0.52475	NA
BERT-Large	0.57491	NA
Covid-Twitter-BERT	0.58712	NA
ConspiracyBERT (BERT-Base)	0.61688	74 hrs
ConspiracyBERT (BERT-Large)	0.63161	125 hrs
ConspiracyBERT (Covid-Twitter-BERT)	**0.63573**	223 hrs

Table 3: Model Results for Multi-label Classification Task on Twitter Conspiracy Dataset

Model	F1 Score	Accuracy Class 0	Accuracy Class 1
TF-IDF + Multinomial Bayes	0.70386	56/283	**705/716**
BERT-Base	0.73199	151/283	646/716
BERT-Large	0.72785	152/283	640/716
Covid-Twitter-BERT	0.76066	156/283	663/716
ConspiracyBERT (BERT-Base)	0.85996	207/283	684/716
ConspiracyBERT (BERT-Large)	0.89678	223/283	696/716
ConspiracyBERT (Covid-Twitter-BERT)	**0.90428**	**225/283**	700/716

Table 4: Fine-tuning results on Conspiracy Stance Dataset

Model	COVID 19 Category (CC)	Stanford Sentiment Treebank v2	SemEval-2016	Average
BERT-Large	0.931	0.937	0.620	0.829
Covid-Twitter-BERT	**0.949**	0.944	0.654	0.849
ConspiracyBERT (BERT-Base)	0.901	0.927	0.595	0.808
ConspiracyBERT (BERT-Large)	0.920	0.941	0.635	0.832
ConspiracyBERT (Covid-Twitter-BERT)	0.943	**0.953**	**0.667**	**0.854**

Table 5: Model results on various Sentiment Analysis Datasets

4 Discussion

4.1 Analysis

As we can see in Table 3, the base models had very low F1 scores compared to after pretraining, with BERT Base having a relative improvement of 17.5%, BERT-Large with 9.86%, and Covid-Twitter-Bert with 8.27%. We speculate that the low F1 scores are due to the fact that there is a significant class imbalance between the 12 classes and that conspiracies were not as distinct as we thought going into the training, which makes this perhaps a sub-optimal training procedure.

In Table 4, we can see the performance of the models on the Conspiracy Stance data set contributed by this work. In this step, we see massive improvements to F1 score and achieve 0.90428 as the best performance, a significant improvement over the base models and classical techniques such as TF-IDF. When comparing the base models effect on the fine-tuning accuracy, we note that BERT-Base has a relative improvement of 17.48% with the pretrained model, BERT-Large has 23.20%, and Covid-Twitter-Bert has 18.88%.

Note that the TF-IDF model correctly classifies less than 20% of the tweets not promoting conspiracy correctly as it is heavily biased by the data imbalance to label tweets as promoting conspiracy (it labels around 94% of tweets as promoting conspiracy despite only 70% of the tweets actually promoting conspiracy). This makes sense as all the tweets in our data set discuss conspiracies and thus have a common vocabulary making it difficult for vocabulary based approaches to differentiate the ones that joke about or have negative things to say about conspiracies and those that promote them. We see that even BERT-Base achieves nearly 3x the performance of the TF-IDF approach in correctly classifying the tweets that do not promote conspiracy with little degradation in predicting those that do. We can see that BERT's robustness to syntax and semantics makes it better fit to differentiate these tweets with a similar vocabulary.

In Table 2, we test our models against the aforementioned data sets from Covid-Twitter-Bert's paper and notice that our pretraining procedure improves performance of the Covid-Twitter-BERT model on both SST and SE, but is not beneficial on CC. We think that this is likely due to the fact that Covid-Twitter-BERT was trained on COVID-19 tweets much closer to the domain of the task (news tweets and non-conspiratorial personal narratives). Our model adds in additional training on COVID-19 conspiracies and misinformation which likely moves the model further from the domain of COVID-19 news/personal narratives or at the very leasts adds a bit more noise around the subject of COVID-19 in particular.

We observe that pretraining on conspiracy related tweets helps in context to both our Conspiracy Stance Dataset, but also in typical sentiment analysis problems compared to other Twitter pretrained methods. This is likely because of the divisive language and strongly opinionated semantics seen in the content of the data, creating a more separable semantic space to learn over; which is supported by our results.

4.2 Limitations and Ethical Considerations

One of the major limitations of our data collection procedure is that we had to rely on tweets with hashtags to insure they were topically relevant to our domain. Perhaps an approach which allowed for keywords like "QAnon" as opposed to hashtags would yield a model more robust to different formulations of conspiracies on Twitter. This issue is also present in our validation set which focuses specifically on tweets containing hashtags relevant to conspiracies, so we do not know how well it would perform on data not relevant to the domain or not containing conspiracy tags. In addition, Twitter has taken a lot of steps to curb conspiracies such as banning search queries for hashtags such as "#DeepState" or "#Obamagate," which limited the amount of pretraining data.

There are also some limitations in the training procedure. We used a multi-class classification task to categorize the conspiracy being discussed by a given tweet. The pretraining likely could have been better guided by Masked Language Modeling as the classes that conspiracies belong to are inherently indistinct as was theorized by the idea of conspiracy belief as a monological belief system. Another limitation in our training procedure is the lack of hyperparameter tuning. Given our lack of computational resources, we simply borrowed the hyperparameter values from COVID-Twitter-BERT, so there is likely room for greater model optimization. Finally, we only validated on classification tasks, so little can be said about ConspiracyBERT's performance on other downstream NLP tasks such as entity recognition.

Finally, we hope that our model will be used by content moderators with discretion. Real conspiracies and corruption do unfold online and we would not want our model to suppress those. So, we advise platforms and moderators to use this model not for automated content removal, but rather automated content flagging for later human review or automated content warning placement.

5 Conclusion

We release ConspiracyBERT, a pretrained language model for conspiracy theories on Twitter. ConspiracyBERT was evaluated on a conspiracy stance data classification task of our own creation along with evaluation on several other classification tasks as established in the Covid-Twitter-BERT paper. ConspiracyBERT significantly outperforms COVID-Twitter-BERT and other generic BERT models on the conspiracy stance detection task and generally outperforms the other models on the tasks set out by COVID-Twitter-BERT.

For future work we hope to evaluate ConspiracyBERT on other downstream tasks outside of classification. We also hope to adopt a multi-task approach focused on sarcasms detection in tweets to better differentiate satirical conspiracy tweets. Additionally, we would like to produce a smaller version of this model with low latency that can be used at scale for content moderation.

References

Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. *Journal of Medical Internet Research*, 22(5):e19458, 2020.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms, 2020.

Karma Allen, 2020. URL https://abcnews.go.com/US/sandy-hook-shooting-conspiracy-theorist-arrested-tormenting-families/story?id=68570486.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

Kate Conger. Twitter takedown targets qanon accounts, Jul 2020. URL https://www.nytimes.com/2020/07/21/technology/twitter-bans-qanon-accounts.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

Karen M Douglas, Robbie M Sutton, and Aleksandra Cichocka. The psychology of conspiracy theories. *Current directions in psychological science*, 26(6):538–542, 2017.

Emilio Ferrara. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, May 2020. ISSN 1396-0466. doi: 10.5210/fm.v25i6.10633. URL http://dx.doi.org/10.5210/fm.v25i6.10633.

Andreas Goreis and Martin Voracek. A systematic review and meta-analysis of psychological research on conspiracy beliefs: Field characteristics, measurement instruments, and associations with personality traits. *Frontiers in Psychology*, 10:205, 2019.

Matthew Haag and Maya Salam. Gunman in 'pizzagate' shooting is sentenced to 4 years in prison, Jun 2017. URL https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html.

Binxuan Huang and Kathleen M. Carley. Disinformation and misinformation on twitter during the novel coronavirus outbreak, 2020.

Colin Klein, Peter Clutton, and Vince Polito. Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in psychology*, 9:189, 2018.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019. ISSN 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892, 2013.

Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020.

Adi Robertson. Reddit has banned the qanon conspiracy subreddit r/greatawakening, Sep 2018. URL https://www.theverge.com/2018/9/12/17851938/reddit-qanon-ban-conspiracy-subreddit-greatawakening.

Viren Swami, Tomas Chamorro-Premuzic, and Adrian Furnham. Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Applied Cognitive Psychology*, 24(6):749–761, 2010.

Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PloS one*, 15(6):e0233879, 2020.

Jill Colvin Zeke Miller and Amanda Seitz. Trump praised the supporters of qanon, a conspiracy theory the fbi says is a domestic terrorism threat, Aug 2020. URL https://www.chicagotribune.com/nation-world/ct-nw-trump-qanon-conspiracy-theory-20200820-m6oeff7wojf77dyeupvl7u6xbu-story.html.