2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic

Viet Duong, Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, USA
vduong@ur.rochester.edu, jluo@cs.rochester.edu

Phu Pham, Tongyu Yang
Goergen Institute for Data Science
University of Rochester
Rochester, USA
{ppham2, tyang20}@u.rochester.edu

Yu Wang
Political Science
University of Rochester
Rochester, USA
ywang176@ur.rochester.edu

***Abstract*—In the United States, the country with the highest confirmed COVID-19 infection cases, a nationwide social distancing protocol has been implemented by the President. Following the closure of the University of Washington on March 7th, more than 1000 colleges and universities in the United States have cancelled in-person classes and campus activities, impacting millions of students. This paper aims to discover the social implications of this unprecedented disruption in our interactive society regarding both the general public and higher education populations by mining people's opinions on social media. We discover several topics embedded in a large number of COVID-19 tweets that represent the most central issues related to the pandemic, which are of great concerns for both college students and the general public. Moreover, we find significant differences between these two groups of Twitter users with respect to the sentiments they expressed towards the COVID-19 issues. To our best knowledge, this is the first social media-based study which focuses on the college student community's demographics and responses to prevalent social issues during a major crisis.**

***Index Terms*—COVID-19, Twitter, College Students , Classification, Sentiment Analysis**

I. Introduction

First detected in Wuhan, China on December 31th, 2019, COVID-19, or the coronavirus, outbreak grew rapidly in scale and severity, and was officially declared as a pandemic on March 11th, 2020[1]. As of April 13th, the World Health Organization (WHO) reported 1,812,734 confirmed cases of COVID-19 worldwide, including 113,675 deaths[2]. Due to the novelty and intractability of the virus, the global community, particularly the elderly and those with underlying medical problems[3], are at a high risk for serious health and safety hazard. However, we suspect that the younger and physically healthier population is just as susceptible, though in a different way, to COVID-19. In order to control the spread of the outbreak, non-pharmaceutical interventions and preventive measures such as social-distancing and self-isolation have been implemented worldwide out of utmost necessity, which has led to the large-scale shutdown of public gathering places. As members of an active working and learning society, people who dedicate most of their daily hours at workplaces and educational institutions are highly vulnerable to the impacts of the closure of these facilities.

This is especially true for college students. The response to the COVID-19 pandemic has brought a sudden disruption in the operations of schools, colleges and universities, influencing more than 1.7 billion students in 192 countries[4]. Located at the epicenter of the pandemic, with 579,005 confirmed cases, including 22,252 deaths[5], the U.S educational system, which is one of the largest in the world, has taken the biggest hit. Beginning with the University of Washington, which closed its campus on March 7th, 2020 and moved classes online for its 50,000 students, many colleges have also immediately responded to this outbreak by cancelling all on-campus activities such as workshops, conferences, and sports, and relocating their in-person classrooms to online platforms. As of April 14th, 2020, more than 124,000 U.S. public and private schools have closed due to the virus, affecting at least 55.1 million students[6]. This transition has introduced multiple challenges for students. The foremost concern is related to how the government and education system handle the pandemic crisis with the study-from-home approach. Past surveys suggested that students experience severe limitation on particular subjects that benefit from physical interaction with the materials, and tend to lose the "pacing mechanism" of scheduled lectures, thus have a higher chance of dropping out than those in traditional settings [1], [2]. As the pandemic unfolded, Sahu [3] hinted at other issues related to the closure of schools, apart from online learning, such as international students' travel and students' mental health. This motivates us to provide a more comprehensive study on the student demographics regarding their primary subjects of concern.

[1]https://cnn.it/37I1fhF
[2]https://bit.ly/3ku4qgH
[3]https://bit.ly/34uwzP7
[4]https://bit.ly/37Gpg91
[5]https://bit.ly/37EFFuv
[6]https://bit.ly/37HO14Q

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

In this study, we attempt to explore the responses to the COVID-19 pandemic by Twitter users, with the focus on the college students. Also, we highlight our findings regarding the college student demographics by characterizing the outstanding differences in their behaviors from the general public. Such insights can be vital for educators and policy-makers to measure the effectiveness of their on-going efforts in the global fight against COVID-19 and the protection of our younger population. In addition, we train classification models to identify the demographics of users who posted tweets associated with COVID-19, as well as extracting the sentiments they inherently expressed in their posts. The models can be used in social media platforms to investigate central social problems, with respect to both their universality and degree of impact, and to draw the community's attention towards high-priority targets for addressing such problems.

Our main contributions are several folds: (1) we approach the social issues related to COVID-19 across different demographics using social media by collecting data from Twitter; (2) we employ topic modeling on a novel Twitter dataset to highlight the topical patterns of the ongoing social media discussions during a major crisis; and (3) we implement a transformer model to achieve new state-of-the-art performance for the Twitter sentiment classification task, which allows insights on social media behaviors to be discovered reliably.

II. Related Work

Our study draws knowledge from the body of research on text mining using data from social media during influential events with topic modeling, including periods amidst the spread of a pandemic ([4], [5]). Previous researches have also attempted to make sentiment classification on social media data using neural networks ([4], [6], [7]). Our approach to sentiment analysis involves the implementation of the RoBERTa transformer model [8], whose state-of-the-art sentiment classification prowess remains largely unused for social studies on Twitter.

Also, recent studies on characterizing the demographics of social media users, along the dimensions such as gender, age, and social class ([9]–[11]), showed that mining fine-grained linguistic patterns from the user's Twitter biography (short self-descriptive text) and posts are proven highly precise for certain attributes when properly constructed. Building upon the discoveries of previous works, we design and evaluate our own college student user classification method. This enables us to identify the two pools of users (college students and general public) among college followers on Twitter for the subsequent comparative analysis.

III. Data Collection and Preprocessing

A. Data Collection

In this study, we limit the user population to those who follow the official Twitter accounts of colleges in the U.S. News 2020 Ranking of Top 200 National Universities. Relevant users identified as English speakers were collected using the Tweepy API (https://git.io/JvAjh). We removed overlapping college followers, as well as extracting their personal information and profile images, to obtain a dataset of 12,407,254 unique users. This set of users is relatively large and 1,641,582 of of these users have Twitter protected accounts, which means we are not allowed to collect tweets (Twitter posts) from them. Thus, we randomly sampled 100,000 users from the unprotected pool to represent the population of college followers for the subsequent tweet collection and text analysis.

Using the Tweepy API, we retrieved a total of 1,873,022 tweets from the 100,000 user samples posted within the timeframe between January 20th, when the first COVID-19 case was confirmed in the U.S., and March 20th of 2020 to cover a two-month period, when nationwide social distancing protocol and school closure were attracting mass concerns. We then extracted tweets related to COVID-19, with a list of keywords consisted of "corona", "#Corona, "#coronavirus", "covid-19", "covid19", "coronavirus", "#Covid_19", "chinese virus", and "#ChineseVirus". As a result, we obtain 73,787 unique COVID-19 related tweets, pertaining to 12,776 users, whom in this study we will address as *affected users*. In addition, tweets that are not related to COVID-19 of the 12,776 *affected users* are kept for the student inference task.

B. Text Preprocessing

We develop a text preprocessing pipeline similar to that of Baziotis et al. [7] to ensure that our text dataset is to a high degree lexically comparable to natural language. This is done by performing sentiment-aware tokenization, spell correction, word normalization, segmentation (for splitting hashtags), and token annotation, using COVID-19 domain-specific word statistics from a novel COVID-19 dataset (https://git.io/JTo1h). Moreover, all texts are lower-cased, while URLs, emails and mentioned usernames are annotated with common designated tags and removed to retain the natural language elements from the text data. The processed tweets are then annotated by the Standford CoreNLP English annotator (https://git.io/JTKZZ), which uses syntactic constituency and dependency tree parsing to extract the appropriate part-of-speech (POS) tags and lemmas (the base/dictionary forms of words) from the tweet tokens.

IV. inference of College student demographics

A. Extracting Age, Gender and Organization Attributes from Twitter User Profiles

We consider age, gender and organization entity to be highly descriptive attributes to first obtain a general view of our user samples. According to National Center for Education Statistics (https://nces.ed.gov/), as of Fall 2017, 56.2% of enrolled students aged between 19 and 29 years old, 20.1% are under 18, and 56.6% of them were female. These student demographic statistics are projected by NCES to remain consistent through 2020. Also, organizational Twitter accounts apparently should not be targeted for student inference because college students are individuals. These attributes are extracted using the M3 (Multilingual, Multimodal, Multi-attribute) deep learning system for inferring the demographics of users from

four sources of information from Twitter profiles: user's name (first and last name in natural language), screen name (Twitter username), biography (short self-descriptive text), and profile image [11]. We extract 1,111 organization entities from 12,776 *affected users*, and disregard them from comparative analysis since they are not individuals. Also, the gender and age attributes are used to verify our classification results.

B. Heuristically Identifying College Students Using Tweets

1) Gold-Standard Annotations: We sample 2,400 random users from the 11,165 non-organizational *affected users* and includes their names, profile images, biographies, and tweets from 1/20 to 3/20/2020. This information is used by human annotators[7] to answer the prompt: "Would you think this person is a COLLEGE STUDENT?" with two response options: "Yes" or "No".

2) Supervised Classification: We encode the standard Bag of N-grams (for 1 up to 4-grams) representation of the user's tweets, which has been highly effective in text categorization and information retrieval [12], to use them as features for our classifiers. To increase the generality of our Bag of N-grams features, we preprocess the tweets as described above and apply TF-IDF vectorization, a term re-weighting scheme that discounts the influence of common terms. We train a Random Forest classifier and report the accuracy: the percentage of correctly labeled users on 20% of the labeled samples. The Random Forest classifier, trained on 1,920 examples, performs quite well with Bag of N-grams features by correctly labeling **78%** of the college students on the test set.

3) Using Heuristic to Override the Classifier: Regarding the self-distinguishing attributes of Twitter users from tweets, Bergsma and Van Durme [9] discovered that users most frequently reveal their attributes in the possessive construction, that is "my X" where X is an attribute, quality or event that they possess (in a linguistic sense). As a matter of fact, we found 306 tweets with the phrase "my class" among the 1,156,947 tweets from non-organizational users. On the contrary, phrases like "I have/had (a) class(es)" occur only 16 times. Therefore, we extract this "my X" attribute type for the college student demographic as follows: we first part-of-speech tag our data using the Stanford CoreNLP tagger and then look for "my X" patterns where X is a sequence of tokens terminating in a noun. To calculate the association between the attributes and the college student demographic, we compute the pointwise mutual information [13] between each attribute A and student over the set of occurrences. If $PMI > 0$, the observed probability of a student and attribute co-occurring is greater than the probability of co-occurrence that we would expect if student and attribute A were independently distributed.

$$PMI(A, student) = \log \frac{p(A, student)}{p(A)p(student)} \quad (1)$$

We employ two techniques for selecting distinctive attributes for college students: (1) we rank the attributes by their PMI scores and use a threshold to select the top-ranked attributes; (2) we manually filter the remaining set of attributes to select those that are judged to be discriminative, including phrases closely associated with college students such as "my zoom class", "my professor", "my dorm", etc. Then we use a simple heuristic to use our identified self-distinguishing attributes in conjunction with a classifier trained on gold-standard annotations: If the user has any self-distinguishing "my-X" attributes, we assign the user to be a college student; otherwise, we trust the output of the classifier. We experiment with our "my-X" attributes and set the PMI threshold to 0.5, and then manually filter out the irrelevant attributes. Applying our heuristics to override the classifier improve the accuracy further to **83%** on the same test set. Therefore, we have firm grounds to utilize the combined classifier and "my-X" heuristics to automatically label college students from the remaining unannotated users, which account for an additional 2,575 out of the total of 3,460 college student users (31% of the non-organizational users).

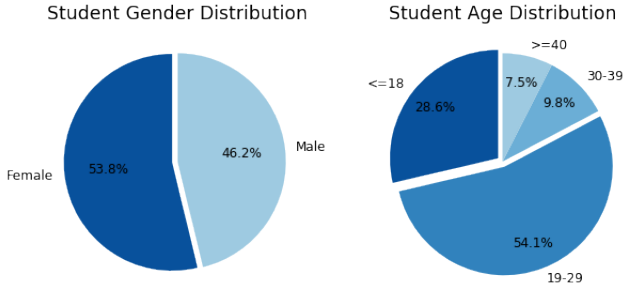

Fig. 1. Gender and Age Distributions of 3,460 College Student Users.

Looking at the age and gender distributions of the college students in our samples (Figure 1), the statistics are very consistent with real world data in the U.S. as 53.8% of the college students we identified are female, which is very close to the 56.7% female percentage predicted by NCES for 2020. Although age classification is a challenging task for the M3 model (0.522 Macro-F1) and even human [11], our results are still within a reasonable margin with NCES's 2020 projection, with 54.1% of the students in the 19-29 age group (vs. 56.7%) and 28.6% under 18 (vs. 21.2%).

V. Topical Analysis of COVID-19 Tweets

To understand the latent topics of the COVID-19 tweets for college followers, we utilize Latent Dirichlet Allocation (LDA) [14] to label universal topics demonstrated by the users. To reduce the complexity of the LDA analysis corpus, only the lemmas of tokens with POS tags of type noun, verb, adjective, and adverb are kept from the preprocessed tweets, because they possess the most meaningful contents related to the topics we are looking to discover. Also, bigrams and trigrams are computed and added to the corpus to account for highly-correlated groups of two and three words. Since certain terms frequently appear in those COVID-19 tweets (e.g. virus, disease, infection, case, test, etc.), we transform our LDA corpus using TF-IDF vectorization. We finetune our

[7]Annotators are IRB certified for Social-Behavioral-Educational Research

LDA model and arrive at the optimal topic number of 55 and coherence score of 0.373.

TABLE I
TOPIC LABELS AND TOP 20 FREQUENCY-WEIGHTED TOPIC KEYWORDS

Topic Label	Subtopic Keywords
Global News	case, confirm, death, report, number, total, someone, disease, break, worldwide, contact, recover, experience, update, state, italy, rise, china, currently, health
Political Discussion	free, vaccine, pay, pass, house, contain, step, point, vote, bill, family, fox, testing, lead, republican, refuse, response, monday, access, senate
Social Distancing	stay, home, watch, safe, worry, india, datum, social_distancing, precaution, sick, work, expect, meeting, practice, project, joke, tour, wash_hand, social
School Closing	school, concern, shut, university, feel, affect, order, cure, student, threat, washington, close, city, closure, problem, campus, knock, imagine, due, follow_government_instruction
Local News	positive, test, county, morning, shit, okay, symptom, find, break, name, isolate, department, resident, case, feel, far, presumptive, health, pm, notice
China Controversy	chinese, call, question, racist, seriously, tonight, guy, employee, blame, take, refer, president, trump, china, perspective, people, guideline, illness, denial_normal, extend

We then label the 6 most frequently discussed topics using the top 20 weighted topic keywords (Table I). Evidently, global news is the most popular topic among the tweets, as the numbers of confirmed positive COVID-19 cases and deaths are constantly increasing globally. The presence of political discussions, as well as the controversy related to the Chinese origin of the virus, is very strong, due to the ongoing presidential election campaign in the US, which gives solid evidence that the COVID-19 pandemic is influencing our political picture. The third and fourth most frequent topics involve social distancing and the closure of colleges.

VI. TOPIC-BASED SENTIMENT ANALYSIS

To expand our study beyond the topic modeling results, we dive deeper into the posts belonging to the each of the 6 most frequently discussed topics. Specifically, for each topic, we separate the college students and general population into two pools and apply the RoBERTa model [8] to classify and examine the sentiments they expressed. We utilize the *transformers* (https://git.io/JfUEh) library by *huggingface* (https://git.io/JToPH), which includes $\text{RoBERTa}_{\text{BASE}}$ in Pytorch [15], and implement the sentiment analysis model with an additional linear layer on top of the pretrained model's outputs. We train and evaluate our models on the SemEval-2017 Task 4A dataset for Twitter message sentiment classification. In the end, our classifier sentiment classifier substantially outperforms the top two performers of SemEval-2017 (Table II) on the test dataset with 0.806 Macro-F1 score. Also, we use the same topic modeling techniques as described in Section V to provide microscopic explanations to the sentiment results.

A. A depressing outlook of COVID-19

Overall, a very small percentage of positive sentiments are expressed towards the COVID-19 tweets (lightest-colored blocks of Figure 2). Also, more than one in five people

TABLE II
PERFORMANCE COMPARISON BETWEEN PREVIOUS METHODS ON TWITTER SENTIMENT CLASSIFICATION AND OURS

Model	Accuracy	Macro-F1 Score
$\text{RoBERTa}_{\text{BASE}}$	**0.806**	**0.806**
LSTMs+CNNs [6]	–	0.685
BiLSTMs+Attention [7]	–	0.677

of our user sample discussed COVID-19 related issues in a negative light. Considering that 2,281 out of a million of the U.S population are physically affected by COVID-19, which is already dangerous, the amount of negativity exhibited on Twitter is very alarming as well. Evidently, not only the COVID-19 pandemic is a health and safety hazard, it also has gloom-ridden impacts on our society. Moreover, for the topic related to the "Chinese virus" controversy, there is an overwhelming number of negative responses. We can see from Table I that "racist" is the fourth most frequent keyword of this topic, which suggests that many of Twitter users associated calling "coronavirus" the "Chinese virus" with racism.

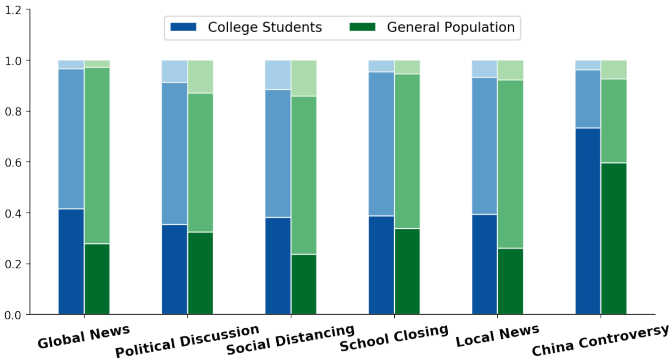

Fig. 2. Sentiment Distributions (%) towards the 6 Most Frequent Topics. Sentiment Percentage Blocks from Bottom to Top: Negative, Neutral, Positive.

B. College students respond more negatively to COVID-19

An important trend that outweighs the rest of our results is that there is a significantly higher percentage among the student population expressing negative sentiments towards the central issues of COVID-19, especially on news related to the spread of the pandemic and social distancing. This shows that our analysis of the data is consistent with the our speculation on the influence of the COVID-19 crisis on our younger population. College students are likely to express negative feelings towards how social distancing and school closure are affecting their work and study environments. Moreover, they tend to be subjected to more negative emotions upon receiving news of the outbreak, which might be due to the subsequent implications of these issues on those more related to their lives.

C. Negativity among College Students via Topic Modeling

We focus on examining the subtopics of School Closing, which is of high concern among college students, and on Social Distancing and China Controversy, where highest gaps in the negative sentiments between college student and the general population are observed (14.5% and 13.8% absolute difference in percentage of negative tweets respectively). Also, only negative and positive tweets are considered because they

provide the most meaningful contexts associated with their sentiments. In general, non-neutral tweets on the Social Distancing and School Closing topics express worrying emotions towards COVID-19, and all the tweets revealing concerns on school closure are negative. Moreover, many students bared aggression to the foreign community, blaming them for the current disruptions in their lives as a result of social distancing. College students also disclosed details of their online learning experience, and mostly showed dislikes for remote study (81.3%). To reflect on the responses of college students on COVID-19 in a more positive light, it is encouraging that our college community remains aware and vocal on the racism problem related to the "Chinese virus" controversy, which sends a powerful message on the public's intolerance of racist behaviors on social media for the betterment of our society.

TABLE III
SUBTOPICS OF SOCIAL DISTANCING

Subtopic Label	**Subtopic Keywords**	**Negative Tweets**
Showing aggression	asia, people, stay, home, worse_european, everyone, piss, fucking, work, fight	81.5%
Detailing precautions	worry, safe, world, tour, stay, people, cancel, take, wash_hand, precaution	65.1%
Expressing concerns	sick, know, go, work, see, watch, really, think, grocery_store, family	85.5%

TABLE IV
SUBTOPICS OF SCHOOL CLOSING

Subtopic Label	**Subtopic Keywords**	**Negative Tweets**
Detailing current situations	knock, follow_government_instruction, feel, survival_rate_whole, country_panicking, get, people, fuck, right, week, campus	98.5%
Detailing remote study	school, close, shut, student, find, get, campus, live_streaming_instead, email, anymore	81.3%
Expressing concerns	people, due, cancel, go, concern, tell, week, imagine, nasty, break	100%

TABLE V
SUBTOPICS OF CHINA CONTROVERSY

Subtopic Label	**Subtopic Keywords**	**Negative Tweets**
Calling out racism	chinese, president, call, racist, refer, flu, reply, people, fuck, trump	95.1%
Addressing attitudes	people, take, perspective, sick, seriously, asian, friend, know, ass, time	97.3%
Detailing public response	call, guy, get, keep, chinese, think, remember, wuhan, response, piss	91.7%

VII. CONCLUSIONS AND FUTURE WORK

We have analyzed 73,787 tweets from 12,776 Twitter college followers who posted tweets related the COVID-19 pandemic, in terms of the outstanding topics on several social issues. We find significant differences in the sentiments expressed towards those topics between the users who are identified as colleges students and those of the general population. Although the percentages of positive COVID-19 tweets are very low for both demographics, college students are shown to be significantly more negative. In addition, microscopic examination of the positive and negative tweets reveals their overwhelmingly troubled feelings amidst the spread of COVID-19, as well as unfavorable reactions to the disruption in their lives such as racism-charged aggression. Moreover, we discover a shift in the target of racism during COVID-19 towards the East Asian community, which the majority of college students and the general public are against.

Since high accuracy is achieved in both of our demographic and sentiment classification models, future studies may collect larger datasets to achieve better performance. In addition, this research mainly focuses on high-level attributes of tweets such as topic models and sentiments in understanding the characteristics of users who discussed social issues associated with COVID-19. Analysis on more fine-grained linguistic information, such as emotion, hate speech, and racism detection can be performed to gain further insights on the more specific COVID-19 related issues detailed in our study.

REFERENCES

[1] L. V. Fedynich, "Teaching beyond the classroom walls: The pros and cons of cyber learning." *J. of Instructional Pedagogies*, vol. 13, 2013.

[2] G. R. Morrison, S. J. Ross, J. R. Morrison, and H. K. Kalman, *Designing effective instruction*. John Wiley & Sons, 2019.

[3] P. Sahu, "Closure of universities due to coronavirus disease 2019 (covid-19): Impact on education and mental health of students and academic staff," *Cureus*, vol. 12, no. 4, 2020.

[4] E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song, "Topic-based content and sentiment analysis of ebola virus on twitter and in the news," *Journal of Information Science*, vol. 42, no. 6, pp. 763–781, 2016.

[5] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs," *arXiv preprint arXiv:2005.03082*, 2020.

[6] M. Cliche, "BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017.

[7] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] S. Bergsma and B. Van Durme, "Using conceptual class attributes to characterize social media users," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 710–720.

[10] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PloS one*, vol. 10, no. 3, 2015.

[11] Z. Wang, S. A. Hale, D. Adelani, P. A. Grabowicz, T. Hartmann, F. Flö"ck, and D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *Proceedings of the 2019 World Wide Web Conference*. ACM, 2019.

[12] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[13] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[14] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 856–864.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.