AMICA: Alleviating Misinformation for Chinese Americans

Xiaoxiao Shang
Department of Computer Science and Engineering
Santa Clara University
Santa Clara, California, USA
xshang@scu.edu

Ye Chen
Department of Computer Science and Engineering
Santa Clara University
Santa Clara, California, USA
ychen40@scu.edu

Yi Fang*
Department of Computer Science and Engineering
Santa Clara University
Santa Clara, California, USA
yfang@scu.edu

Yuhong Liu*
Department of Computer Science and Engineering
Santa Clara University
Santa Clara, California, USA
yhliu@scu.edu

Subramaniam Vincent*
Markkula Center for Applied Ethics
Santa Clara University
Santa Clara, California, USA
svincent@scu.edu

ABSTRACT

The increasing popularity of social media promotes the proliferation of misinformation, especially in the communities of Chinese-speaking diasporas, which has caused significant negative societal impacts. In addition, most of the existing efforts on misinformation mitigation have focused on English and other western languages, which makes numerous overseas Chinese a very vulnerable population to online disinformation campaigns. In this paper, we present AMICA, an information retrieval system for alleviating misinformation for Chinese Americans. AMICA dynamically collects data from popular social media platforms for Chinese Americans, including WeChat, Twitter, YouTube, and Chinese forums. The data are stored and indexed in Elasticsearch to provide advanced search functionalities. Given a user query, the ranking of social media posts considers both topical relevance and the likelihood of being misinformation.

CCS CONCEPTS

• **Information systems → Information extraction**; • **Computing methodologies → Natural language processing**.

KEYWORDS

Chinese misinformation, Social media, Information retrieval

ACM Reference Format:
Xiaoxiao Shang, Ye Chen, Yi Fang, Yuhong Liu, and Subramaniam Vincent. 2023. AMICA: Alleviating Misinformation for Chinese Americans. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591810

1 INTRODUCTION

Chinese Americans, a rapidly growing group in the US electorate, rely heavily on information in their native language. However, the Chinese media sphere in the US has been neglected, and questionable content has gained popularity among Chinese-speaking immigrants. Concerns have been raised about disinformation and polarization campaigns targeting US-based Chinese language speakers [10]. The distribution and consumption of misinformation are further complicated by differences in values, interests, and communication systems.

In this paper, we present AMICA, an information retrieval system for **A**lleviating **MI**sinformation for **C**hinese **A**mericans. AMICA is a repository and monitor that stores and ranks Chinese language posts from social media and web forums, making it easier to search for disinformation and misinformation. The data is indexed in Elasticsearch for advanced search functionalities. The ranking of social media posts considers both relevance and likelihood of misinformation when responding to user queries. AMICA can aid newsrooms, journalists, and anti-disinformation initiatives in capturing emerging disinformation topics and propagation patterns, and prepare for future reporting. The system can be used as labeled training data for research efforts on disinformation. To the best of our knowledge, there exists no prior work on developing an information retrieval system for combating misinformation for overseas Chinese.

2 RELATED WORK

Over the past few years, the Chinese-language based misinformation propagation in online social media, including topics on COVID, Ukraine war, the political tensions between U.S. and China, etc., has attracted extensive attention [1, 2, 5, 17]. In [6], the authors observed 18 Chinese-based propaganda techniques. Although some of these techniques are propaganda-specific, there are some techniques that can be effective for general misinformation propagation and are adopted in this study. In [19], the authors identify that WeChat, a Chinese-language based social media platform, provides

[*] Corresponding authors

SIGIR '23, July 23–27, 2023, Taipei, Taiwan.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591810

key clues to how misinformation is constructed and distributed among the Chinese-American community.

Many existing studies have revealed the unique propagation patterns of misinformation in online social media [3, 11, 16, 20], which can serve as effective indicators for learning schemes to detect misinformation. However, very few studies are exploring misinformation propagation patterns in the Chinese-language context. In addition, due to the challenge of data collection, there are limited data sets containing fine-grained temporal information to reflect the dynamic evolution of propagation data. In this study, we adopt a dynamic data crawling approach, which can well balance the trade-off between data collection costs and granularity of the propagation data.

Beyond academia research, there are ongoing efforts from practitioners. For example, Google's Fact Check Explorer[1] is a search engines where users can search for potential misinformation articles, while it does not support Chinese language. Piyaoba[2] is a non-profit fact-checking portal to combat misinformation and disinformation directed at the Chinese American community, and they rely on domain experts to manually identify and debunk misinformation posts from social media.

3 SYSTEM DESIGN

AMICA is a web-based search engine that allows users to search for information by entering keywords or phrases. It returns a list of relevant results based on the user's query. The results are ranked by taking into account various factors such as relevance, sentiment, and propagation data. Its backend includes crawlers to collect data, Elasticsearch[3] for storage, and two ranking scores (topical relevance and machine learning-based misinformation likelihood) are combined to present results to the user. The major components and pipeline of AMICA are shown in Figure 1.

3.1 User Interface

The frontend of AMICA is a web search page built with Flask [9], a lightweight Python web framework, that allows users to search for information on it. The user interface of the web search page is built using HTML, CSS, and JavaScript, and can be customized to provide a user-friendly experience.

AMICA has a secure login system that allows authorized users to access the search engine and conduct searches. Once logged in, users can enter search queries into a search box, view results on the page, and interact with the results in several ways, including clicking on links and filtering results based on other criteria. Search results typically include the essential information about the post, such as the post headline, author, first published date, etc. To improve the user experience, each search result includes a preview on the page. Screenshots of the frontend are included in Section 4.

3.2 Data Scraping

AMICA integrates social media data from Twitter, YouTube, WeChat, and Chinese forums. We automatically crawl the data from Twitter and YouTube with their official APIs and from WeChat with an open-source program. We collect articles, posts, and transcripts from monitored accounts and regularly obtain propagation data, such as reads and views, to analyze misinformation spread. Monitored accounts are identified manually and from reports by Chinese-language fact-checking organizations [14]. Our data collection program uses four crawler instances and stores data in Elasticsearch for convenient querying and modification by APIs.

Twitter: We collect data using the Twitter official API [18], which regularly monitors target accounts, captures newly posted data, and stores them in the proposed repository. Data includes basic account and tweet information, as well as propagation data such as likes, retweets, and comments, with collection frequency and duration to be discussed at the end of this section.

YouTube: We utilize Google developer accounts and the YouTube API V3 [8] to collect basic information about YouTube accounts and videos, including user name, user id, video id, title, description, URL, transcript (if available), and published date, as well as propagation data similar to the Twitter data, and their corresponding time stamps, following the settings of our dynamic crawling mechanism.

WeChat: Being a "closed" system, WeChat lacks an official API for data collection. Thus we use an open source program called WeChat Spider [12] that captures communication data via a proxy between the server and user app, which requires logging into WeChat on a phone or emulator and using JavaScript code to collect data on targeted articles or profile pages. Due to WeChat's detection and blocking of crawling behavior, data collection is limited in speed, frequency, and duration. Therefore, we mainly focus on collecting basic information such as account name, published articles, dates, headlines, full text, and the number of reads with timestamps.

Bulletin-based Chinese Forums: We collect data from Chinese bulletin-based forums, such as WenXue City[4] and Huaren.us[5], using our own web crawler based on Selenium [15], a web crawling tool. The collected data includes account name, article headline, full text, read number, replies number, URL, and published date.

Dynamic Crawling: We balance accuracy and collection costs by carefully designing the frequency of data crawling for propagation data (e.g., likes, retweets), due to limited API calls allowed by social media platforms. We observe a bursty human dynamics pattern where most propagation occurs within 24 hours of post publication, with few changes after 72 hours. Thus, we use a dynamic crawling window, with high frequency in the first 24 hours and low frequency from 24 to 72 hours.

3.3 Text Pre-processing

To enable further analysis, the posts undergo text normalization, tokenization, and part-of-speech (POS) tagging to assign labels such as verbs or nouns. In this stage, stopwords are also removed to improve the efficiency of the process and reduce noise in the data. To facilitate the two ranking systems employed in AMICA, different analyzers are used in this process as described below.

The Smart Chinese Analysis plugin (*smartcn*)[6] is used in Elasticsearch built-in ranking system to enhance the search engine's

[1] https://toolbox.google.com/factcheck/explorer
[2] https://www.piyaoba.org/
[3] https://github.com/elasticsearch/elasticsearch
[4] https://www.wenxuecity.com/
[5] https://huaren.us/home/index
[6] https://www.elastic.co/guide/en/elasticsearch/plugins/current/analysis-smartcn.html

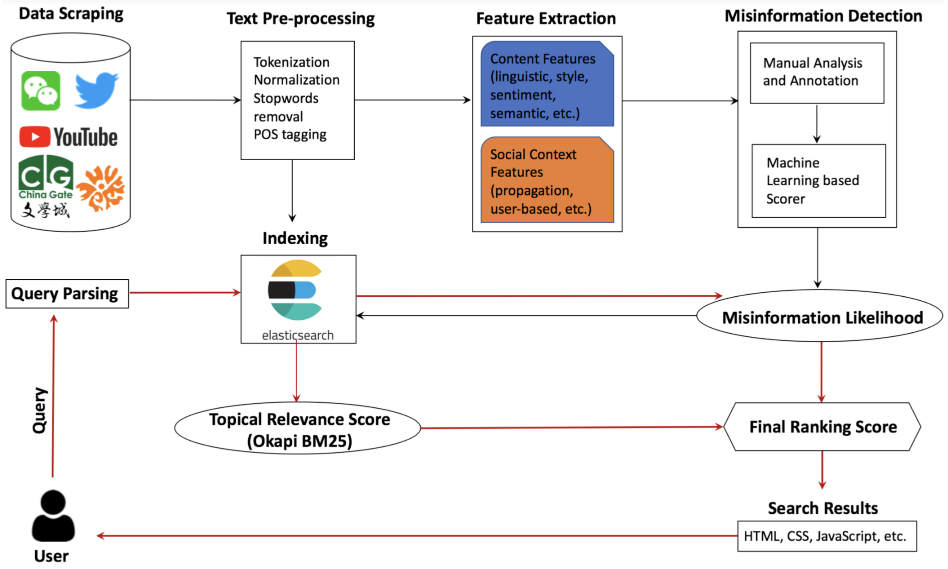

Data Scraping
YouTube
China Gate
文學城
Text Pre-processing
Tokenization
Normalization
Stopwords removal
POS tagging
Feature Extraction
Content Features (linguistic, style, sentiment, semantic, etc.)
Social Context Features (propagation, user-based, etc.)
Misinformation Detection
Manual Analysis and Annotation
Machine Learning based Scorer
Indexing
elasticsearch
Query Parsing
Misinformation Likelihood
Query
Topical Relevance Score (Okapi BM25)
Final Ranking Score
Search Results
HTML, CSS, JavaScript, etc.
User

Figure 1: The pipeline of AMICA with the major components. The black arrows represent offline computation before a query is given and the red arrows indicate online processing when a user query is issued at run time.

performance on Chinese texts. This plugin is designed specifically for the Chinese language and improves the engine's understanding of Chinese words, resulting in more accurate and relevant search results for users.

For Chinese text segmentation, *Jieba*[7], a Chinese text analysis package in Python is utilized, which uses a dictionary-based approach to tokenize Chinese texts into words and phrases. It also supports part-of-speech tagging and keyword extraction. After text pre-processing the text data, Elasticsearch indexes each post according to a mapping property that has been predefined (i.e. Okapi BM25). This mapping property specifies the structure of the data and how it will be stored and searched, allowing for fast and efficient retrieval of information when queried.

3.4 Feature Extraction

To determine if a post contains misinformation, the following features are extracted by AMICA:

Content Features: which measures textual characteristics of the content, including 1) the min-max normalize length of the post; 2) the number of question marks, as multiple questions in a post, often used in user posts invoking worry or concern; 3) the number of exclamation marks, as many exclamation marks are used for attention-grabbing language, often seen in user generated content posts seeking attention; 4) strong attempts to gather attention by using language that stokes fear, anxiety, worry, or shock, often seen in posts with exclamation marks in quick succession. Examples (translated) include *Shocking!*, *Disappointed!*, *OMG!*, *Warning!*, *WHAT!*, *Rarely seen in history!*, etc.; 5) the sentiment of a post, which can be determined using natural language processing techniques. The package *cntext* [7] is leveraged to perform Chinese text analysis using traditional methods such as word count and sentiment analysis. In order to extract attention-gathering features, a pre-defined dictionary containing commonly used attention-gathering words is utilized.

Social Context Features: which measures social characteristics of the post, including 1) where the post was originally published; 2) how the post has spread or propagated, such as the number of likes, shares/retweets, or comments, and how quickly it has spread; 3) sharing evidence or references to support claims by citing other sources.

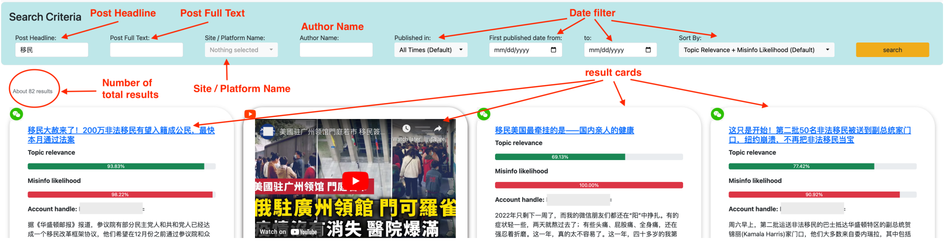
3.5 Misinformation Detection

To measure the likelihood of a given post being misinformation, we deploy a machine learning approach. Specifically, we used the logistic regression model [4] by taking the features discussed in the previous section as input and generating a score by the Sigmoid function indicating the probability that the content is misinformation. To create training data for the model, two domain experts manually labeled posts from the social media platforms that we monitored. We randomly sampled posts from the entire corpus and treated them as non-misinformation since most of the posts in the corpus are legitimate content. In the ongoing work, we are utilizing the known misinformation posts manually identified by some Chinese fact-checking websites such as Piyaoba as part of the training data. We are also experimenting with more advanced machine learning models as more training data are accumulated. It is worth noting that the misinformation likelihood is query independent. Once the model is trained, we calculate the misinformation likelihood for all the posts in the corpus offline and store them in the Elasticsearch database, which will be later combined with the topical relevance to generate the final ranking given a query.

3.6 Ranking

For a given query, the final ranking score of a post combines topical relevance (query-dependent) with misinformation likelihood

[7] https://github.com/fxsjy/jieba



Post Headline: 移民

Post Full Text:

Site / Platform Name: Nothing selected

Author Name:

Published in: All Times (Default)

First published date from: mm/dd/yyyy

to: mm/dd/yyyy

Sort By: Topic Relevance + Misinfo Likelihood (Default)

About 82 results

移民大赦来了！200万非法移民有望入籍成公民，最快本月通过法案

Topic relevance 93.83%

Misinfo likelihood 98.22%

Account handle: [redacted]

据《华盛顿邮报》报道，参议院有部分民主党人和共和党人已经达成一个移民改革框架协议，他们希望在12月份之前通过参议院和众

美國駐广州领馆門庭若市 移民簽...

美國駐广州领馆 門庭

俄駐廣州領館 門可羅雀

疫情沒有消失 醫院爆滿

移民美国最牵挂的是——国内亲人的健康

Topic relevance 69.13%

Misinfo likelihood 100.00%

Account handle: [redacted]

2022年只剩下一周了，而我的微信朋友们都还在"阳"中挣扎。有的症状轻一些，两天就熬过去了；有些头痛、屁股痛、全身痛，还在强忍着折磨。这一年，真的太不容易了。这一年，四十多岁的我第

这只是开始！第二批50名非法移民被送到副总统家门口，纽约崩溃，不再把非法移民当宝

Topic relevance 77.42%

Misinfo likelihood 90.92%

Account handle: [redacted]

周六早上，第二批运送非法移民的巴士抵达华盛顿特区的副总统贺锦丽(Kamala Harris)家门口，他们大多数来自委内瑞拉，其中包括

Figure 2: Screenshot of the AMICA result page with results based on the search criteria with the query "移民 (immigrants)" in the post headline.

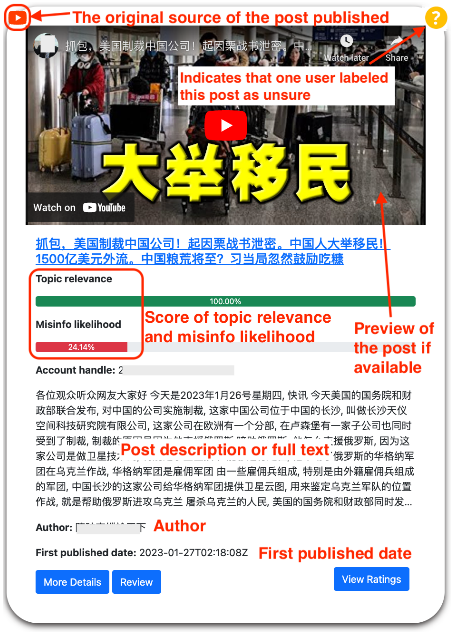
(query-independent) introduced in Section 3.5. We used the Elasticsearch built-in Okapi BM25 [13] to measure the topical relevance, factoring in term frequency and document length. The posts on different social media platforms may have different characteristics. For example, tweets on Twitter are usually much shorter than an article on WeChat, which may result in BM25 scores of different ranges across platforms. To make all the scores (both in BM25 and misinformation likelihood) comparable, we use min-max normalization to normalize the scores within each respective social platform. The posts/articles are then ranked based on the descending order of the sum of BM25 and misinformation likelihood scores. The frontend will render the page based on this result.

4 DEMONSTRATION

In this section, we briefly demonstrate the main functionalities of AMICA. As illustrated in Figure 2, the search criteria bar is located at the top of the page allows users to enter their queries and search based on various parameters including post headline, post full text, site/platform name, and author name. Users can further refine their search results based on the publication date of each post. The search results are displayed using Bootstrap[8] cards, and the total number of results is indicated above the result cards.

Each result card shown in Figure 3 contains basic information about a post, as well as its topic relevance score and misinformation likelihood, with a logo on the top left corner indicating the original platform or media where the post was published. In addition, there is a button for more details that opens a pop-up window displaying a line chart for propagation data, showing how the post has spread over time, and other information about the post. There is also a button for human review, which allows users to submit a review form to determine whether the post is misinformation or not. The review will be stored in an Elasticsearch database, and each user can only submit one review per post. This feature helps improve the accuracy of the misinformation likelihood score, as it enables human experts to verify whether a post is actually misinformation or not. We also have a button that allows users to view other users' reviews in the database for the same post. Due to the space limit, we skip the screenshot for those three buttons.

As of this writing, AMICA is monitoring 119 accounts on Twitter (85), WeChat (11), and YouTube (23), with a total of 163,930 records in our database. Twitter has the highest number of records at 158,240, followed by WeChat with 3,065 and YouTube with 2,625. The numbers are growing every day as data are being constantly collected.


The original source of the post published
?
Indicates that one user labeled this post as unsure
大举移民
Watch on YouTube
Watch later
Share
Score of topic relevance and misinfo likelihood
Preview of the post if available
Post description or full text
Author
First published date
More Details
Review
View Ratings

抓包，美国制裁中国公司！起因栗战书泄密。中...

抓包，美国制裁中国公司！起因栗战书泄密。中国人大举移民！1500亿美元外流。中国粮荒将至？习当局忽然鼓励吃糠

Topic relevance 100.00%

Misinfo likelihood 24.14%

Account handle: 2[redacted]

各位观众听众网友大家好 今天是2023年1月26号星期四, 快讯 今天美国的国务院和财政部联合发布, 对中国的公司实施制裁, 这家中国公司位于中国的长沙, 叫做长沙天仪空间科技研究院有限公司, 这家公司在欧洲有一个分部, 在卢森堡有一家子公司也同时受到了制裁, 制裁的 援俄罗斯, 因为这家公司是做卫星技 俄罗斯的华格纳军团在乌克兰作战, 华格纳军团是雇佣军团 由一些雇佣兵组成, 特别是由外籍雇佣兵组成的军团, 中国长沙的这家公司给华格纳军团提供卫星云图, 用来鉴定乌克兰军队的位置作战, 就是帮助俄罗斯进攻乌克兰 屠杀乌克兰的人民, 美国的国务院和财政部同时发...

Author: [redacted]下

First published date: 2023-01-27T02:18:08Z

Figure 3: Screenshot of one example result card.

5 CONCLUSION AND FUTURE WORK

AMICA is a search engine designed to counter misinformation. It includes a backend system that utilizes a web crawler, database, and machine learning model to score data and provide users with relevant search results. The frontend is a web search page built with Flask and features a secure login system for authorized users. AMICA is currently being tested with real-world users, and we are continuously improving it based on user feedback. As data accumulate, we plan to leverage advanced IR/NLP/ML techniques to improve accuracy and scalability of the system.

ACKNOWLEDGMENTS

We appreciate the valuable feedback of the reviewers. This research was in part supported by the Miami Foundation and the Ciocca Center at Santa Clara University. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors. We thank Piyaoba for helpful discussions and providing us with a list of social media accounts of interest to monitor. We also thank Ruoxi Chi, Jil Patel, and Jinming Nian for their work on manual analysis and frontend development.

[8] https://getbootstrap.com/

REFERENCES

[1] April, 2022. *Information manipulation on Ukraine, Shanghai Outbreak, Transnational Repression in the US*. https://freedomhouse.org/report/china-media-bulletin/2022/information-manipulation-ukraine-shanghai-outbreak-transnational

[2] Vivian Wang Amy Qin and Danny Hakim. Nov. 2020. *How Steve Bannon and a Chinese Billionaire Created a Right-Wing Coronavirus Media Sensation*. https://www.nytimes.com/2020/11/20/business/media/steve-bannon-china.html

[3] Mauro Barni, Yi Fang, Yuhong Liu, Laura Robinson, Kazutoshi Sasahara, Subramaniam Vincent, Xinchao Wang, Zhizheng Wu, et al. 2022. Combating Misinformation/Disinformation in Online Social Media: A Multidisciplinary View. *APSIPA Transactions on Signal and Information Processing* 11, 2 (2022).

[4] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning* (1st ed.). Springer, Chapter 4.3.2 Logistic Regression, 163–168.

[5] Ariel Bogle and Iris Zhao. October, 2020. *Anti-Beijing group with links to Steve Bannon spreading COVID-19 misinformation in Australia*. https://www.abc.net.au/news/science/2020-10-09/anti-beijing-group-with-links-to-steve-bannon-misinformation/12735638

[6] Rong-Ching Chang, Chun-Ming Lai, Kai-Lai Chang, and Chu-Hsing Lin. 2021. Dataset of Propaganda Techniques of the State-Sponsored Information Operation of the People’s Republic of China. *arXiv preprint arXiv:2106.07544* (2021).

[7] Xudong Deng and Peng Nan. 2022. cntext: a Python tool for text mining. https://doi.org/10.5281/zenodo.7063523

[8] Inc. Google. 2013. Youtube API v3. https://developers.google.com/youtube/v3.

[9] Miguel Grinberg. 2018. *Flask web development: developing web applications with python*. " O’Reilly Media, Inc.".

[10] Scott W Harold, Nathan Beauchamp-Mustafaga, and Jeffrey W Hornung. 2021. *Chinese Disinformation Efforts on Social Media*. Technical Report. RAND PROJECT AIR FORCE SANTA MONICA CA.

[11] V Indu and Sabu M Thampi. 2021. A systematic review on the influence of User personality in rumor and misinformation propagation through social networks. In *Advances in Signal Processing and Intelligent Recognition Systems: 6th International Symposium, SIRS 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 6*. Springer, 216–242.

[12] Qiang Li. 2018. *Wechat Spider*. https://github.com/lqqyt2423/wechat_spider

[13] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends® in Information Retrieval* 13, 1 (2018), 1–126. https://doi.org/10.1561/1500000061

[14] Piyaoba.org. 2022. 2022 Q3 Piyaoba Disinformation Report (Full English Version). https://www.piyaoba.org/wp-content/uploads/2022/12/2022-Q3-Piyaoba-Disinformation-Report-Full-English-Version.pdf Accessed: 2022-09-22.

[15] Sujay Raghavendra and Sujay Raghavendra. 2021. Introduction to selenium. *Python Testing with Selenium: Learn to Implement Different Testing Techniques Using the Selenium WebDriver* (2021), 1–14.

[16] Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. 2022. Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)* 16, 3 (2022), 1–34.

[17] Olivia Solon. June, 2022. *Chinese Government Asked TikTok for Stealth Propaganda Account*. https://www.bloomberg.com/news/articles/2022-07-29/chinese-government-asked-tiktok-for-stealth-propaganda-account?cmpid=socialflow-twitter-business#xj4y7vzkg

[18] Inc. Twitter. 2020. Twitter API v2. https://developer.twitter.com/en/docs/twitter-api.

[19] Chi Zhang. 2018. *WeChatting American Politics: Misinformation, Polarization, and Immigrant Chinese Media*. https://www.cjr.org/tow_center_reports/wechatting-american-politics-misinformation-polarization-and-immigrant-chinese-media.php#takeaways

[20] Qiang Zhang, Jonathan Cook, and Emine Yilmaz. 2021. Detecting and forecasting misinformation via temporal and geometric propagation patterns. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 455–462.