AROT-COV23: A DATASET OF 500K ORIGINAL ARABIC TWEETS ON COVID-19

Cheng Xu
School of Computer Science
University College Dublin
Dublin, Ireland
cheng.xu1@ucdconnect.ie

Nan Yan
Brooklyn College
The City University of New York
New York, USA
nan.yan80@bcmail.cuny.edu

ABSTRACT

This paper presents a dataset called AROT-COV23 (**AR**abic **O**riginal **T**weets on **COV**ID-19 as of 20**23**) containing about 500,000 original Arabic COVID-19-related tweets from January 2020 to January 2023. The dataset has been analyzed using a corpus-based approach to identify common themes and trends in the data and gain insights into the ways in which Arabic Twitter users have discussed the pandemic. The results of the analysis are also presented and discussed in terms of their implications for the field of Natural Language Processing (NLP) in Africa and for understanding the role of Twitter in the spread of COVID-19-related information in the region.

1 INTRODUCTION

The COVID-19 pandemic Organization (2020) has had a significant impact on the global population, and has generated a large volume of online discourse related to the virus and its consequences Cinelli et al. (2020); Goel & Gupta (2020); Tsao et al. (2021). In Africa, social media platforms such as Twitter have played a particularly important role in the spread of COVID-19-related information, as they provide a means of communication that is both fast and accessible to many users Obi-Ani et al. (2020); Olatunji et al. (2020).

There are several reasons why Arabic might be a good choice for building a corpus of COVID-19-related text in Africa. First, Arabic is a widely spoken language in the region, so there is likely to be a large volume of Arabic COVID-19-related text available for collection Lodhi (1993). Second, Arabic is an important language for international communication, so there may be a significant amount of Arabic COVID-19-related information available online, including news articles, official statements, and social media posts Versteegh (2014).

Given the importance of Twitter in the spread of COVID-19-related information in Africa, there is a need for tools and techniques that can help to analyze and understand the content of Arabic COVID-19-related tweets Alsudias & Rayson (2020); Essam et al. (2021). Natural language processing (NLP) techniques, which allow computers to analyze and understand human language, can be particularly useful in this context. However, the development of NLP resources and tools for African languages has often been overlooked, leading to a lack of support for many African languages in the field Blodgett & O'Connor (2017).

In this paper, we present a dataset called AROT-COV23[1] (**AR**abic **O**riginal **T**weets on **COV**ID-19 as of 20**23**) with about 500,000 original Arabic COVID-19-related tweets from January 2020 to January 2023, and demonstrate how this dataset can be used to gain insights into the content and sentiment of Arabic COVID-19 tweets. By analyzing the corpus using a corpus-based approach, we are able to identify common themes and trends in the data, and shed light on the ways in which Arabic Twitter users have discussed the pandemic.

The remainder of this paper is organized as follows. In Section 2, we describe the process of collecting and pre-processing the data for our dataset. In Section 3, we present the results of our analysis, including an overview of the common themes and sentiment expressed in the data. In Section 4, we

[1] The dataset is available at https://github.com/chengxuphd/AROT-COV23

discuss the implications of our findings for the field of NLP in Africa and for understanding the role of Twitter in the spread of COVID-19-related information in the region. Finally, in Section 5, we conclude the paper and suggest directions for future research on NLP in Africa.

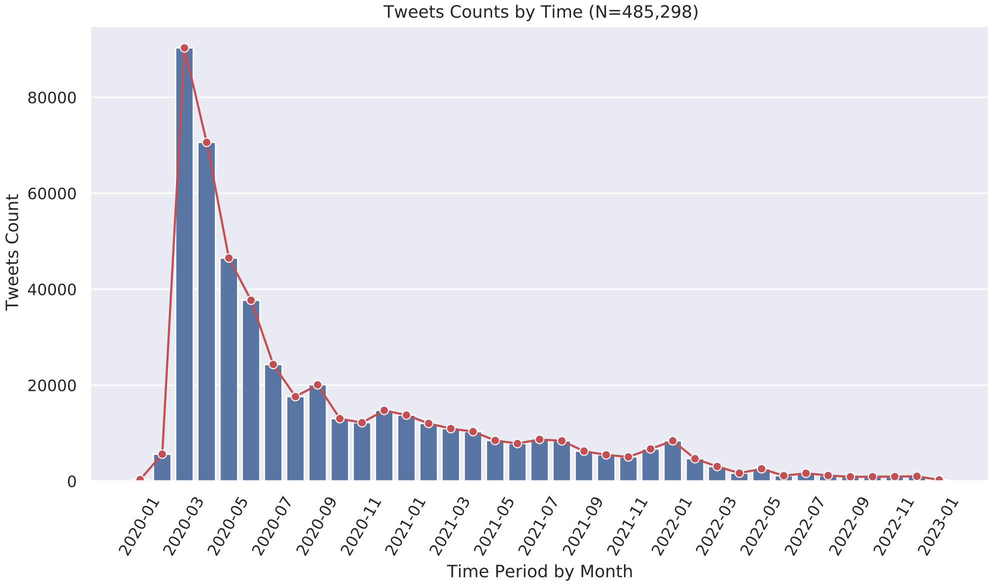

Figure 1: Temporal distribution of data in the AROT-COV23 by month. The data range from January 1, 2020 to January 5, 2023, with 485,298 data items, and concentrated in the first half of 2020, peaking at 90,301 in March and then falling.

2 Related Work

While African languages have historically received less attention in NLP research, the fact remains that Africa is home to over a billion people and over two thousand languages Abo et al. (2019); Tubishat et al. (2019). As such, it is important to consider the perspectives and needs of African language speakers in the field of NLP. By mainstreaming African languages in NLP research, we can ensure that these voices are heard and valued in the broader research community. Through this analysis of Arabic COVID-19 tweets, we aim to make a contribution to the growing body of NLP research on African languages. We hope that this work will help to increase the visibility and impact of NLP research in African languages within the broader research community.

There is a substantial amount of text in Arabic that is available for NLP research, but not much research has been done specifically on the topic of COVID-19. In this section, we present a review of a selection of relevant works in this area. Our review provides an overview of the current state of the field and highlights key areas for future research.

Alsudias & Rayson (2020) analyzed the tweets in three different ways: identifying the topics discussed, detecting rumors, and predicting the source of the tweets. This work is representative work, but it does not analyze the sentiment trends in the corpus. In the Haouari et al. (2021) and Alqurashi et al. (2020), the authors present a dataset of Arabic COVID-19 tweets, includes approximately 2.7 and 3.9 million tweets respectively. While these work provides a description and statistical overview of the dataset and suggests some potential application scenarios, it does not delve more deeply into the data to conduct a more in-depth analysis. Ali (2021) collected online learning-related tweets in Arabic and conducted a comprehensive emotion mining and sentiment analysis (SA) during the COVID-19 pandemic. They then employed Information Gain (IG) as a filtering technique and identified the latent reasons behind negative sentiments. The experiments showed that the proposed model had a maximum accuracy of around 89.6% using the SVM classifier. Addawood (2020) collected 3.8 million tweets using hashtags and keywords. Then carried on the simple statistics to the

انوروك سوريف
ةحصلا ةرازو
ديدجلا انوروك
دجتسملا انوروك
ديفوك انوروك
ةيملاعلا ةحصلا

Figure 2: The word cloud generated based on the AROT-COV23 dataset.

database. Bahja et al. (2020) and Elsaka et al. (2022) used machine learning techniques to detect the sentiment of Arabic tweets, such as Safety, Worry, and Irony.

In conclusion, the field of NLP research on African languages, particularly Arabic, has seen a growing interest in recent years, particularly in the context of the COVID-19 pandemic. However, there is still a need for more in-depth and comprehensive analysis of the text data available in these languages. The studies reviewed have primarily focused on detecting rumors, predicting the source of tweets and also, but there is a need for more advanced analysis, such as sentiment trend analysis. Additionally, while datasets of Arabic COVID-19 tweets have been made available, there is a need for more detailed analysis of these datasets to uncover insights and trends.

3 Dataset Description

The AROT-COV23 dataset is a large-scale collection of original Arabic tweets related to COVID-19, spanning from January 2020 to January 2023, and the period for which we collected the data runs from January 1, 2020 to January 5, 2023, with more details in Figure 1. The dataset contains approximately 500,000 original tweets, providing a rich source of information on how Arabic-speaking Twitter users have discussed and shared information about the pandemic.

The AROT-COV23 dataset we proposed was assembled utilizing the Twitter API[2]. Due to the restrictions imposed by Twitter's Developer Agreement and Policy on Content redistribution[3], the data that we make available does not comprise direct tweet text data. It is to be noted that all open-source data is intended for non-commercial research purposes exclusively; for further information, please refer to the GitHub page associated with this dataset. In the collection and pre-processing section of the dataset, we made a more detailed description in Appendix A.

We hope that the AROT-COV23 dataset will contribute to the research community's efforts to better understand the impact of COVID-19 on the Arabic-speaking world and to develop more effective strategies for managing pandemics in the future.

4 Analysis and Potential Tasks

In addition to collating the Arabic tweet data into AROT-COV23, we also did some simple analysis on this dataset, and provide some potential tasks that can be used for this dataset. One of the first

[2] https://developer.twitter.com/en/docs/twitter-api

[3] https://developer.twitter.com/en/developer-terms/agreement-and-policy

analyses we conducted was word cloud analysis after pre-processing all tweets, and the results of the result are shown in Figure 2. The word with the highest frequency is فيروس كورونا, which means ”Corona Virus” in English. The word cloud analysis gave us an overview of the most common words used in the tweets, and it served as a starting point to identify the main topics discussed in the tweets. Based on the results, it is clear that all tweets are related to COVID-19.

Subsequently, we also used a Bert-based sentiment analysis model for Arabic, called CAMeLBERT-DA SA Model Inoue et al. (2021), which was built by fine-tuning the CAMeLBERT Dialectal Arabic (DA) model. The final prediction results are in Figure 3. As can be seen from the figure, the sentiment information of most of the data is on the neutral side (61.545%), followed by the negative side (29.631%), and the least amount of positive sentiment information (8.824%). It is important to note that while the results obtained through the sentiment analysis model may contain some inaccuracies, they provide an overall representation of the sentiment information distribution within the entirety of the dataset. Additionally, it is worth mentioning that the dataset comprises solely of original tweets, providing a reflection of mainstream opinions to a certain degree and reducing the presence of extreme emotions. As such, it is logical to deduce that tweets with neutral sentiment comprise a significant proportion of the dataset.

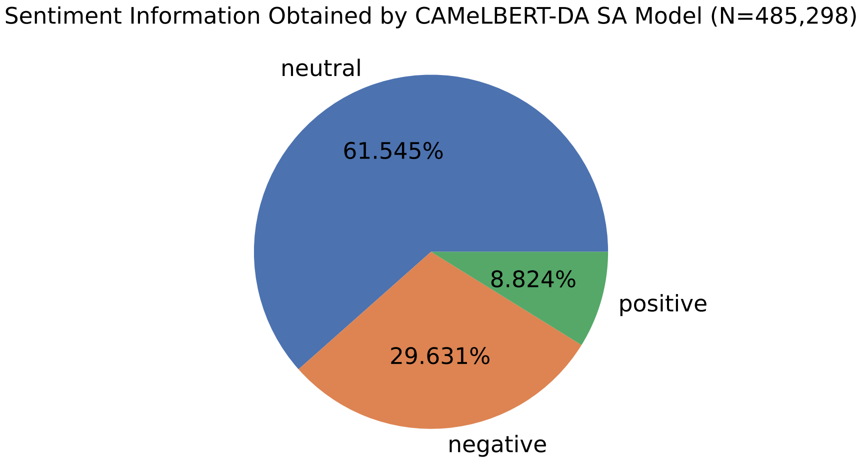

Figure 3: The results of predicting the sentiment information of AROT-COV23 using the CAMeLBERT-DA SA Model.

There are a variety of potential tasks that can be performed using the AROT-COV23 dataset. In addition to word frequency statistics and tasks related to sentiment information, based on the AROT-COV23 dataset, we believe there are several potential tasks as follows.

1. Network analysis Cheng et al. (2021); Vosoughi et al. (2018): visualizing the interactions and relationships between tweet authors, such as who is retweeting or replying to whom, and identifying key influencers or opinion leaders.

2. Stance detection AlDayel & Magdy (2021); Hanselowski et al. (2018): identifying the stance or opinion of the tweet authors towards the pandemic, such as whether they are supportive or critical of government actions.

3. Cross-lingual analysis Zhang et al. (2021); McCarthy et al. (2019): comparing the discourse and sentiment of COVID-19 tweets in Arabic with tweets in other languages to identify similarities and differences in how the pandemic is discussed in different cultures and regions. For example, comparing the sentiment differences between English-speaking, Chinese-speaking and Arabic-speaking social groups on the COVID-19.

4. Topic modeling Wallach (2006); Boon-Itt et al. (2020): identifying common themes or topics discussed in the tweets, such as information about COVID-19 symptoms, vaccine rollout, or government response to the pandemic.

5. Various scenarios requiring the Arabic corpus and various unsupervised learning tasks, such as fine-tune on this database for unsupervised fake news detection Sheng et al. (2022); Wu et al. (2021), and the characteristics of this dataset were used to model the spread of the disease to discover similar patterns Xu et al. (2022; 2021).

5 Conclusion

In this paper, we presented the AROT-COV23 dataset, which contains about 500,000 original Arabic COVID-19-related tweets from January 2020 to January 2023. The dataset has been analyzed using a corpus-based approach to identify common themes and trends in the data and gain insights into the ways in which Arabic Twitter users have discussed the pandemic. The results of the analysis have been discussed in terms of their implications for the field of NLP in Africa and for understanding the role of Twitter in the spread of COVID-19-related information in the region. The paper suggests that this dataset can be used for various NLP tasks and that mainstreaming African languages in NLP research is important to ensure that the perspectives and needs of African language speakers are heard and valued in the broader research community. This work aims to increase the visibility and impact of NLP research in African languages within the broader research community.

Author Contributions

All authors contributed to the design of the study. NY collected and preprocessed the data, CX performed the analysis and wrote the manuscript.

Acknowledgments

This work acknowledges the data provided by Twitter, Inc.

References

Mohamed Elhag Mohamed Abo, Ram Gopal Raj, and Atika Qazi. A review on arabic sentiment analysis: state-of-the-art, taxonomy and open research challenges. *IEEE Access*, 7:162008–162024, 2019.

Aseel Addawood. Coronavirus: Public arabic twitter data set, 2020. URL https://openreview.net/forum?id=ZxjFAfD0pSy.

Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

Manal Mostafa Ali. Arabic sentiment analysis about online learning to mitigate covid-19. *Journal of Intelligent Systems*, 30(1):524–540, 2021.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*, 2020.

Lama Alsudias and Paul Rayson. COVID-19 and arabic twitter: How can arab world governments and public health organizations learn from social media? In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*, 2020.

Mohammed Bahja, Rawad Hammad, and Mohammed Amin Kuhail. Capturing public concerns about coronavirus using arabic tweets: An nlp-driven approach. In *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*, pp. 310–315, 2020.

Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*, 2017.

Sakun Boon-Itt, Yukolpat Skunkan, et al. Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4):e21978, 2020.

Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 148–157, 2021.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10, 2020.

Tarek Elsaka, Imad Afyouni, Ibrahim Hashem, and Zaher Al Aghbari. Spatio-temporal sentiment mining of covid-19 arabic social media. *ISPRS International Journal of Geo-Information*, 11(9), 2022.

Nader Essam, Abdullah M Moussa, Khaled M Elsayed, Sherif Abdou, Mohsen Rashwan, Shaheen Khatoon, Md Maruf Hasan, Amna Asif, and Majed A Alshamari. Location analysis for arabic covid-19 twitter data using enhanced dialect identification models. *Applied Sciences*, 11(23): 11328, 2021.

Ashish Goel and Latika Gupta. Social media in the times of covid-19. *Journal of clinical rheumatology*, 26(6):220–223, 2020.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1859–1874, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pp. 82–91, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online), April 2021. Association for Computational Linguistics.

Abdulaziz Y Lodhi. The language situation in africa today. *Nordic Journal of African Studies*, 2(1): 11–11, 1993.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 229–244, Florence, Italy, August 2019. Association for Computational Linguistics.

Ngozika A Obi-Ani, Chinenye Anikwenze, and Mathias Chukwudi Isiani. Social media and the covid-19 pandemic: Observations from nigeria. *Cogent arts & humanities*, 7(1):1799483, 2020.

Olusoji S Olatunji, Olusola Ayandele, Doyin Ashirudeen, Oluwatosin S Olaniru, et al. “infodemic” in a pandemic: Covid-19 conspiracy theories in an african country. *Social Health and Behavior*, 3(4):152, 2020.

World Health Organization. Coronavirus disease 2019 (covid-19): situation report, 83. Technical documents, World Health Organization, 2020.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4543–4556, 2022.

Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A Butt. What social media told us in the time of covid-19: a scoping review. *The Lancet Digital Health*, 3(3):e175–e194, 2021.

Mohammad Tubishat, Mohammad AM Abushariah, Norisma Idris, and Ibrahim Aljarah. Improved whale optimization algorithm for feature selection in arabic sentiment analysis. *Applied Intelligence*, 49(5):1688–1707, 2019.

Kees Versteegh. *Arabic language*. Edinburgh University Press, 2014.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359 (6380):1146–1151, 2018.

Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pp. 977–984, 2006.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2560–2569, 2021.

Cheng Xu, Qingling Chen, Fan Ye, Qi Fan, and Qing Wang. Selection of surgical procedures and analysis of prognostic factors in patients with primary gastric tumour based on cox regression: a seer database analysis based on data mining. *Gastroenterology Review/Przeglad Gastroenterologiczny*, 16(2):144–154, 2021.

Cheng Xu, Jing Wang, Tianlong Zheng, Yue Cao, and Fan Ye. Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Archives of Medical Science*, 18(5):1208–1220, 2022.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9220–9230, 2021.

A Collection and Pre-processing of AROT-COV23 dataset.

We employed the use of the Academic Research access level[4] of the Twitter API for the data collection step of our AROT-COV23, which is the official way that Twitter provides researchers with access to tweets for research purposes. We identified our data collection target as original tweets in Arabic pertaining to the COVID-19 pandemic, so the request for the API was set to `lang:ar -is:nullcast -is:retweet -is:reply`, which implicate to find Arabic tweets and filter out marketing tweets, retweets and replies.

The tweets in the AROT-COV23 dataset were collected using a set of COVID-19-related keywords in Arabic. The tweets were then filtered to ensure that they were written in Arabic. We selected three keywords related to COVID-19 for the data request, the details are in Table 1. We set the data fields to be collected and their meanings are presented in Table 2. At the same time, we also randomly selected a tweet data and displayed it in Table 3.

Since the data obtained by requesting Twitter API is in JSON format, we convert it into a standard CSV data file. Then the tweet text data is clear and segmented for subsequent training or prediction, including the removal of punctuation, numbers, emoji, links, etc. Refer to the Github repository[5] of the AROT-COV23 dataset for details.

[4] https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

[5] https://github.com/chengxuphd/AROT-COV23/tree/main/data_collection

Table 1: Three keywords for the data request.

Keyword	**Description**
COVID-19	Coronavirus disease 2019 (COVID-19) was released by the World Health Organization as the most common name for this disease.
مرض فيروس كورونا	This is Arabic for ”Coronavirus Disease” (COVID-19)
فيروس كوفيد	This is Arabic for ”Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)”

Table 2: Tweets field feature information.

Field	**Type**	**Description**
tweet id	string	The unique identifier of the requested Tweet.
author id	string	The unique identifier of this user.
created_at	date	Creation time of the Tweet.
lang	string	Language of the Tweet, if detected by Twitter.
like_count	int	The number of likes on this tweet
quote_count	int	The number of times this tweet has been quoted.
reply_count	int	The number of replies to this tweet.
retweet_count	int	The number of retweets to this tweet.
tweet	string	The actual UTF-8 text of the Tweet.
user_verified	boolean	Indicates if this user is a verified Twitter User.
followers_count	int	The number of followers of the author.
following_count	int	The number of following of the author.
tweet_count	int	Total number of tweets by the author.
listed_count	int	The number of public lists that this user is a member of.
name	string	The name of the user.
username	string	The Twitter screen name, handle, or alias.
user_created_at	date	The UTC datetime that the user account was created
description	string	The text of this user’s profile description (bio).

Table 3: An example of a randomly selected display from the AROT-COV23 database.

Field	Value
tweet id	1233338555252006918
author id	805692634127736832
created_at	2020-02-28 10:30:00+00:00
lang	ar
like_count	25
quote_count	1
reply_count	0
retweet_count	4
tweet	...في الصور الملتقطة ٢٧ فبراير ٢٠٢٠، ٢ من المرضى
user_verified	True
followers_count	667414
following_count	7
tweet_count	121945
listed_count	671
name	CGTN Arabic
username	cgtnarabic
user_created_at	2016-12-05T08:37:47.000Z
description	...شبكة تلفزيون الصين الدولية مؤسسة إعلامية فريدة