A Weibo Dataset for the 2022 Russo-Ukrainian Crisis

Yi R. Fung, Heng Ji
University of Illinois at Urbana-Champaign
{yifung2,hengji}@illinois.edu

Abstract

Online social networks such as Twitter and Weibo play an important role in how people stay informed and exchange reactions. Each crisis encompasses a new opportunity to study the portability of models for various tasks (*e.g.*, information extraction, complex event understanding, misinformation detection, etc.), due to differences in domain, entities, and event types. We present the **Russia-Ukraine Crisis Weibo (RUW) dataset**, with over 3.5M user posts and comments in the first release[1].

1 Introduction

The Russo-Ukrainian conflict stemmed from complex issues that added up since the disintegration of the Soviet Union era, including Ukraine's geo-political divide between its internal East and West (Karácsonyi et al., 2014), NATO eastward expansion (Sarotte, 2014), and Russian interference with Ukraine's eastern breakaway regions (Kofman et al., 2017). The conflict escalated on February 24, 2022, with Russian forces entering Ukraine territory under the campaign of a special military operation (Ellyatt, 2022). At the time of writing, the conflict is, unfortunately, still ongoing. People across the world have been following the news and sharing opinions regarding the situation online.

The study of information discourse and propagation in online social networks is of particular interest to the research community (Sacco and Bossio, 2015). Deeper analysis from information extraction (Wen et al., 2021), argumentation mining (Lawrence and Reed, 2020; Reddy et al., 2021), opinion mining (Ravi and Ravi, 2015), propagation pattern monitoring (Shu et al., 2020), and misinformation detection (Shu et al., 2019; Fung et al., 2021), etc., can help society better understand and manage crises events. To faciliate timely analysis, we thus publish an open dataset from Weibo posts relevant to the ongoing 2022 Russo-Ukrainian crisis, with continued daily updates.

2 Related Work

Various previous attempts have explored online social networks, in the aspects of misinformation detection (Islam et al., 2020), propaganda identification (Khanday et al., 2021), and general network characteristics during crises (Kim and Hastak, 2018). However, news media and news dissemination tend to exhibit different properties under different time and places, leading to a diachronic bias, which limits the portability of existing models (Murayama et al., 2021). While there exists a recent Russo-Ukrainian Crisis Twitter dataset (Haq et al., 2022), no such dataset exists yet for Weibo, the second largest news-oriented social media platform in the world. It is generally interesting to uncover patterns universal across platforms, regulated by a separate set of rules and participated by a separate set of users with different user behavior. Note that Twitter data (multilingual) involves English as the primary language, and Weibo data involves Chinese as the primary language.

3 Dataset Collection

We use the publicly available `weibo-scraper`[2] library package to collect Weibo data on the Russian-Ukraine crisis, including the following information for each public user post.

i User information: profile ID, profile name, profile image, and profile description..

ii Post content: the main text, as well as any images and videos attached, if available.

iii Post metadata: the # of likes and shares on the post.

[1] Our data is available at https://github.com/yrf1/RussiaUkraine_weibo_dataset

[2] https://pypi.org/project/weibo-scraper/

iv Post comments: information similar to i-iii for all comments underneath a user post.

Note, the weibo-scraper library supports querying Weibo tweets by user profile. Hence, we manually prepared a list of the top 120 public user profiles, who have actively posted about and ranked amongst the top posts of trending hashtags related to the Russian-Ukraine crisis. The average number of reactions and shares from the posts of these users are 4338 and 163, respectively. Then, we performed a one-hop expansion to include all the users who commented underneath the post of these active profiles as well.

Since the crisis is still ongoing, the list of user profiles may likely change over the time, with inclusion of new active user profiles according to the situation. We initiated the data collection process on the final week of Feb 2022. As the development of events preceding a conflict breakout is also of interest, we retrospectively collected data within the last three months prior to the onset of the conflict as well. We will continue the data collection and update the data repository accordingly.

4 Dataset Summary

By March 8th, we have collected over 27,341 Weibo posts, and over 3.5 million corresponding Weibo comments from 107,797 unique users, containing keywords relevant to the Russian-Ukraine crisis. In Table 1, we show the Weibo post distribution based on the keywords in this Russia-Ukraine Crisis Weibo (RUW) dataset. Most of the Weibo posts contain the 乌克兰 (Ukraine) and 俄罗斯 (Russia) reference, followed by 俄乌 (Russia-Ukraine). The two countries' leaders, 普京 (Putin) and 泽连斯基 (Zelensky), are mentioned less.

Keyword	# Weibo Posts
乌克兰 (Ukraine)	120436
俄罗斯 (Russia)	95845
俄乌 (Russia-Ukraine)	53164
基辅 (Kiev)	28748
普京 (Putin)	21005
泽连斯基 (Zelensky)	18109
俄方 (Russian side)	10612
北约 (NATO)	10025
乌方 (Ukrainian side)	7181

Table 1: Number of Weibo posts for each keyword, in descending order.

In addition, we show a word cloud of text from the Weibo posts in Figure 1. Besides the named entities (*i.e.*, country name and person name), the prominence of words such as '局势 (situation)' and "制裁 (sanction)' suggests that most of tweets discuss the latest updates on the Russo-Ukrainian crisis.

Figure 1: Word cloud of all Russo-Ukranian Weibo posts.

5 Dataset Usage

There are many interesting task settings that can be experimented with, using our Russo-Ukranian Crises Weibo dataset. These include event clustering, false rumor detection, and evaluating the portability of news analytic methodologies across Twitter and Weibo domains.

6 Ethical Considerations

We recognize that societal response in crises events may typically be polarized or contentious across different geo-political and socio-cultural regions. We urge researchers and analysts to be as objective as possible in their studies of the matter.

References

Holly Ellyatt. 2022. Russian forces invade ukraine. *NBC*.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Tristan Braud, and Pan Hui. 2022. Twitter dataset for 2022 russo-ukrainian crisis.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):1–20.

Dávid Karácsonyi, Károly Kocsis, Katalin Kovály, József Molnár, and László Póti. 2014. East-west dichotomy and political conflict in ukraine-was huntington right? *Hungarian Geographical Bulletin*, 63(2):99.

Akib Mohi Ud Din Khanday, Qamar Rayees Khan, and Syed Tanzeel Rabani. 2021. Identifying propaganda from online social networks during covid-19 using machine learning techniques. *International Journal of Information Technology*, 13(1):115–122.

Jooho Kim and Makarand Hastak. 2018. Social network analysis: Characteristics of online social networks after a disaster. *International journal of information management*, 38(1):86–96.

Michael Kofman, Katya Migacheva, Brian Nichiporuk, Andrew Radin, Jenny Oberholtzer, et al. 2017. *Lessons from Russia's operations in Crimea and Eastern Ukraine*. Rand Corporation.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. 2021. Mitigation of diachronic bias in fake news detection dataset. *arXiv preprint arXiv:2108.12601*.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Vittoria Sacco and Diana Bossio. 2015. Using social media in the news reportage of war & conflict: Opportunities and challenges. *The journal of media innovations*, 2(1):59–76.

Mary Elise Sarotte. 2014. A broken promise: What the west really told moscow about nato expansion. *Foreign Aff.*, 93:90.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, et al. 2021. Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143.