

Social Media Mining for Health 2022 (#SMM4H)
Gyeongju, Republic of Korea

COVID-19-related Nepali Tweets Classification in a Low Resource Setting

Rabin Adhikari, Safal Thapaliya, Nirajan Basnet, Samip Poudel,
Aman Shakya, Bishesh Khanal

Context

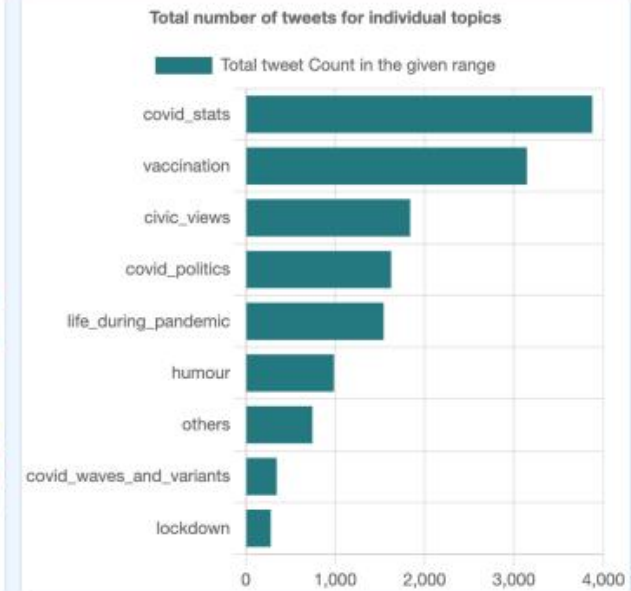
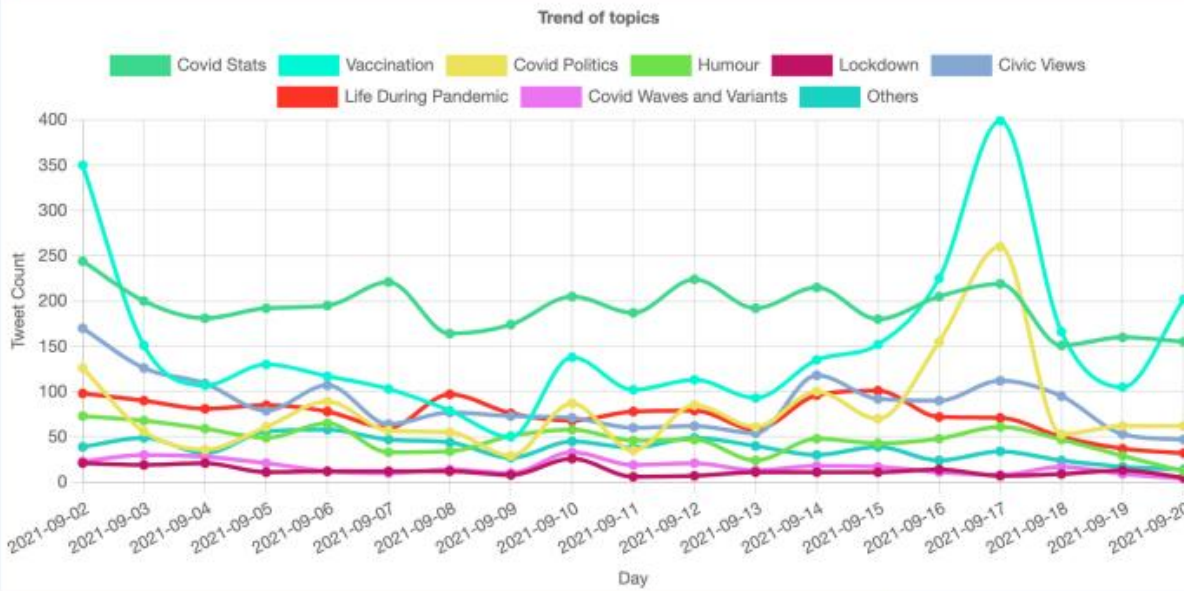
- **Massive use of social media platforms** in local languages during the COVID.
- Organizations like WHO have **developed automated social media analysis tools**.

Problem

- Limited to a very few languages, and several countries are unable to take their benefit.
- Low resource language specific tools have limited coverage.
- Gap in data availability and development of NLP tools for Nepali Language.

Contributions

- **Nepali Annotated Tweets with COVID-19 Topics Classification (NAT-CTC) dataset**
- An **open-source web-based dashboard** for topic analysis with online learning features
- Comparison of **generic vs language family-specific multilingual language models**



Filter by Topic:



Jagdishor Panday on September 18, 2021

परराष्ट्र सचिव भरतराज पौड्यालले कम विकसित मुलुकहरूको मन्त्रीस्तरीय भर्चुअल बैठकमा शुक्रबार खोपको असमान वितरणबारे बोल्दै नेपालसहित कम विकसित मुलुकमा न्यायोचित र सर्वसुलभ रूपमा खोप उपलब्ध गराइनुपर्ने बताए ।

vaccination covid_politics

EDIT

दहाल ऋषी(जय श्रीमन्नारायण) on September 18, 2021

कोभिड ले दुई साल देखी रोक्यो

life_during_pandemic

EDIT

Krishna Gyawali on September 19, 2021

Trending Words



Outline

- NAT-CTC Dataset
- Nepali Tweets Classification
- Results
- Final Remarks

Outline

- **NAT-CTC Dataset**
- Nepali Tweets Classification
- Results
- Final Remarks

Dataset Overview

- Multi-label multi-annotator dataset
- 12,241 tweets in Devanagari script
- Manually tagged with 8 topics by 7 annotators

Keyword-based filtering

- Identified **48 keywords** covering majority of COVID-19-related tweets in Devnagari script
- Used *twarc* [1] to collect tweets tagged as Nepali by Twitter
- New keywords were added iteratively to an initial set of keywords using manual review

Keywords in Nepali	English Translation
कोभिड, कोरोना, कोरोनाभाइरस, कोरोनाभाईरस, कोवीड, कोभीड, कोविड, भ्याक्सिन, भ्याक्सीन, बेड, “अक्सिजन सिलिन्डर”, लकडाउन, निशेधाज्ञा, संक्रमण, संक्रमित, खोप, फाइजर, फाईजर, भेरोसेल, भेरोशेल, “जोर बिजोर”, जोर-बिजोर, “स्वास्थ्य तथा जनसंख्या मन्त्रालय”, “होम आइसोलेसन”, आइसोलेसन, आईसोलेसन, “दोस्रो लहर”, प्रभावकारी, प्रभावकारिता, महामारी, माहामारी, भेरियन्ट, भेरीयन्ट, डेल्टा, “डेल्टा प्लस”, “संक्रमण मुक्त”, मास्क, स्यानिटाइजर, “भौतिक दुरी”, डोज, म्यूटेसन, शय्या, जोखिम, आइसियु, भेन्टिलेटर, एन्टिजेन, कोभ्याक्स, “विश्व स्वास्थ्य संगठन”	COVID, Corona, Corona Virus, Vaccine, Bed, Oxygen Cylinder, Lockdown, Infection, Infected, Fizer, Verocel, Home Isolation, Isolation, Second Wave, Ministry of Health and Population, Efficient, Efficacy, Pandemic, Epidemic, Variant, Delta, Delta Plus, Infection-free, Mask, Sanitizer, Social Distancing, Dose, Bed, Mutation, ICU, Ventilator, Antigen, COVAX, World Health Organization

[1] <https://github.com/DocNow/twarc>

8 COVID-19-related Topics

- Referred to the 30 topics used in WHO EARS [1]
- Contextualized and developed **8 topics suitable to describe specific narratives** in Nepal

COVID Stats

Vaccination

COVID Politics

Humor

Lockdown

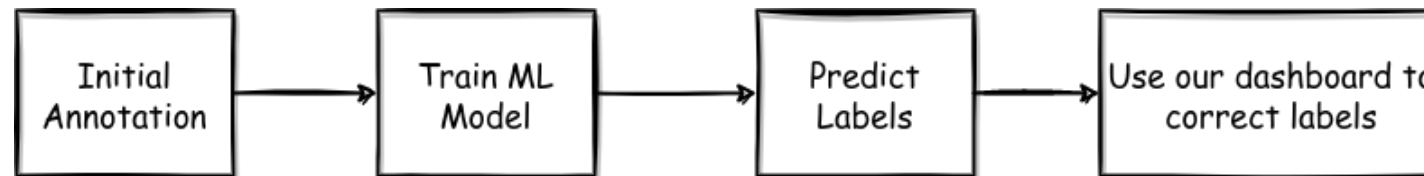
Civic Views

Life during
Pandemic

Waves and
Variants

Annotation with Incremental Learning

- 7 annotators used *Label Studio* [1] to annotate the tweets

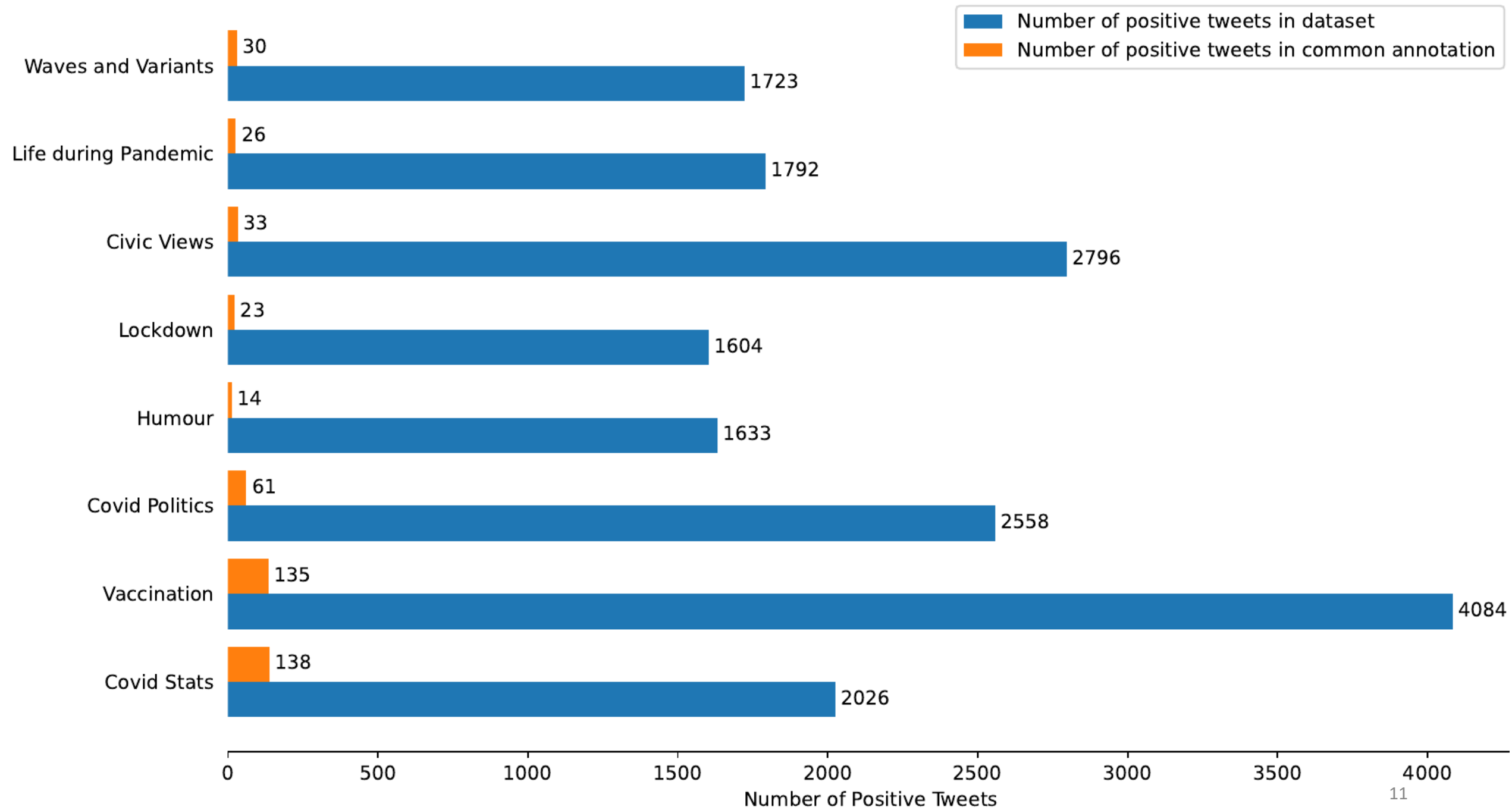


- This process **increased the annotation speed** and **improved the ML model**
- **400 tweets** were used to study **inter-rater agreement** between **4 annotators**
- Mean Fleiss' **Kappa [2,3] score of 0.64**

[1] <https://labelstud.io/>

[2] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

[3] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.



Outline

- NAT-CTC Dataset
- **Nepali Tweets Classification**
- Results
- Final Remarks

Tweets Preprocessing

User mentions and link removal

Remove blank spaces


Remove texts followed by "via"

Lowercase Latin characters

Removal of tweets with length ≤ 3

Normalize Unicode strings to NFKC standard

कोरोना संक्रमणको दर बढ्यो ~~https://t.co/zx0inwXmpr~~ via ~~@himal_dainik~~ 

छिटो खोप लगाऔं ~~https://t.co/qveQjvKg0x~~ 

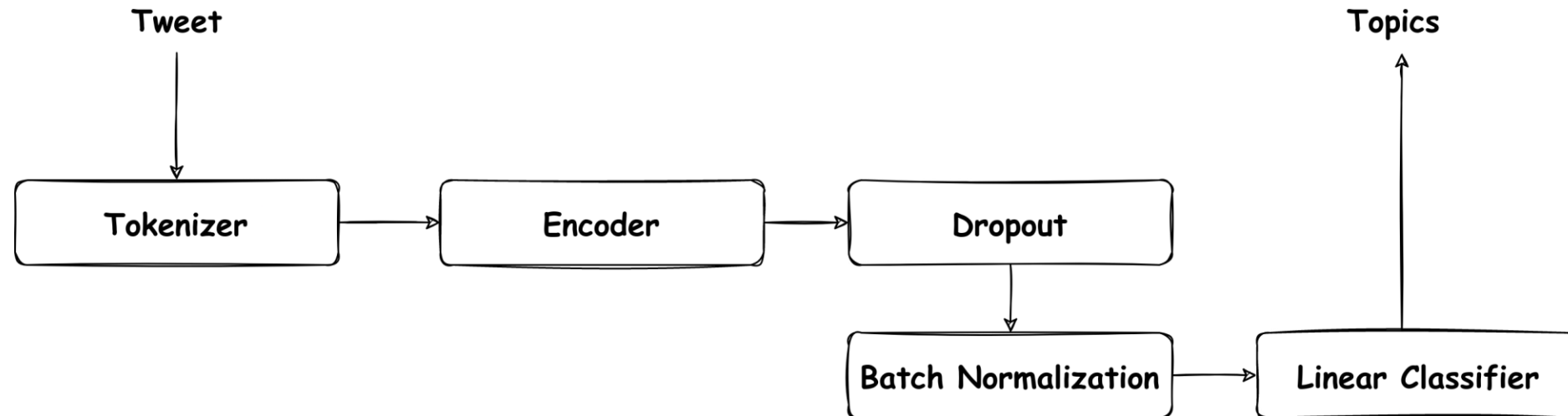
Multilingual Language Models

1. **mBERT:** Generic with more Latin Scripts ~ 104 languages [1]
2. **MuRIL:** More emphasis on Indic Languages ~ 17 languages [2]

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[2] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuriL: Multi-lingual representations for indian languages. ArXiv preprint arXiv:2103.10730.

Model Pipeline



Outline

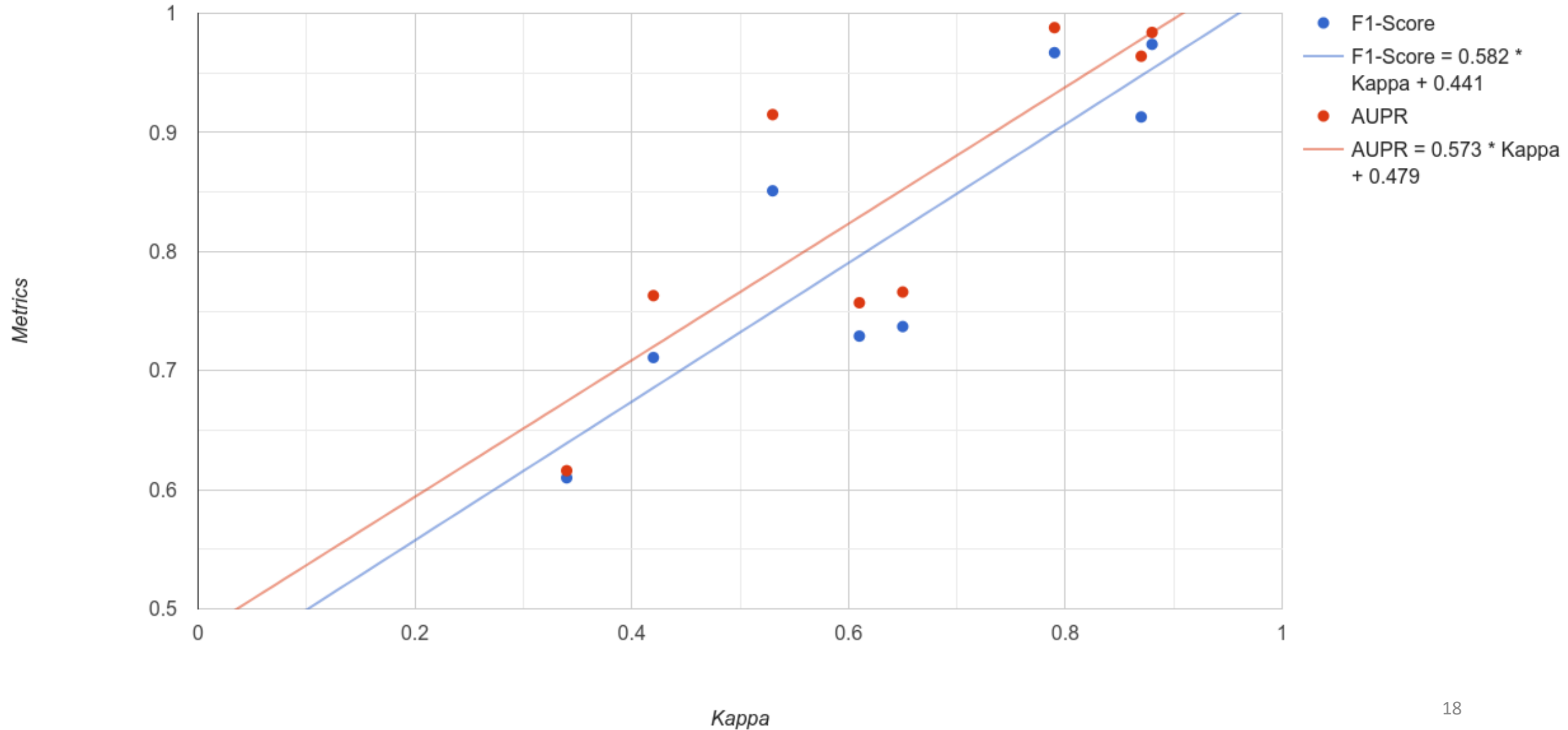
- NAT-CTC Dataset
- Nepali Tweets Classification
- **Results**
- Final Remarks

Topic-wise Model Performance

Topics	F1-Score	Area under PR Curve
COVID Stats	0.91 ± 0.01	0.96 ± 0.00
Vaccination	0.97 ± 0.00	0.98 ± 0.00
COVID Politics	0.71 ± 0.01	0.76 ± 0.01
Humor	0.74 ± 0.01	0.77 ± 0.01
Lockdown	0.97 ± 0.01	0.99 ± 0.00
Civic Views	0.73 ± 0.01	0.76 ± 0.01
<i>Life during Pandemic</i>	<i>0.61 ± 0.03</i>	<i>0.62 ± 0.04</i>
Waves and Variants	0.85 ± 0.01	0.92 ± 0.01

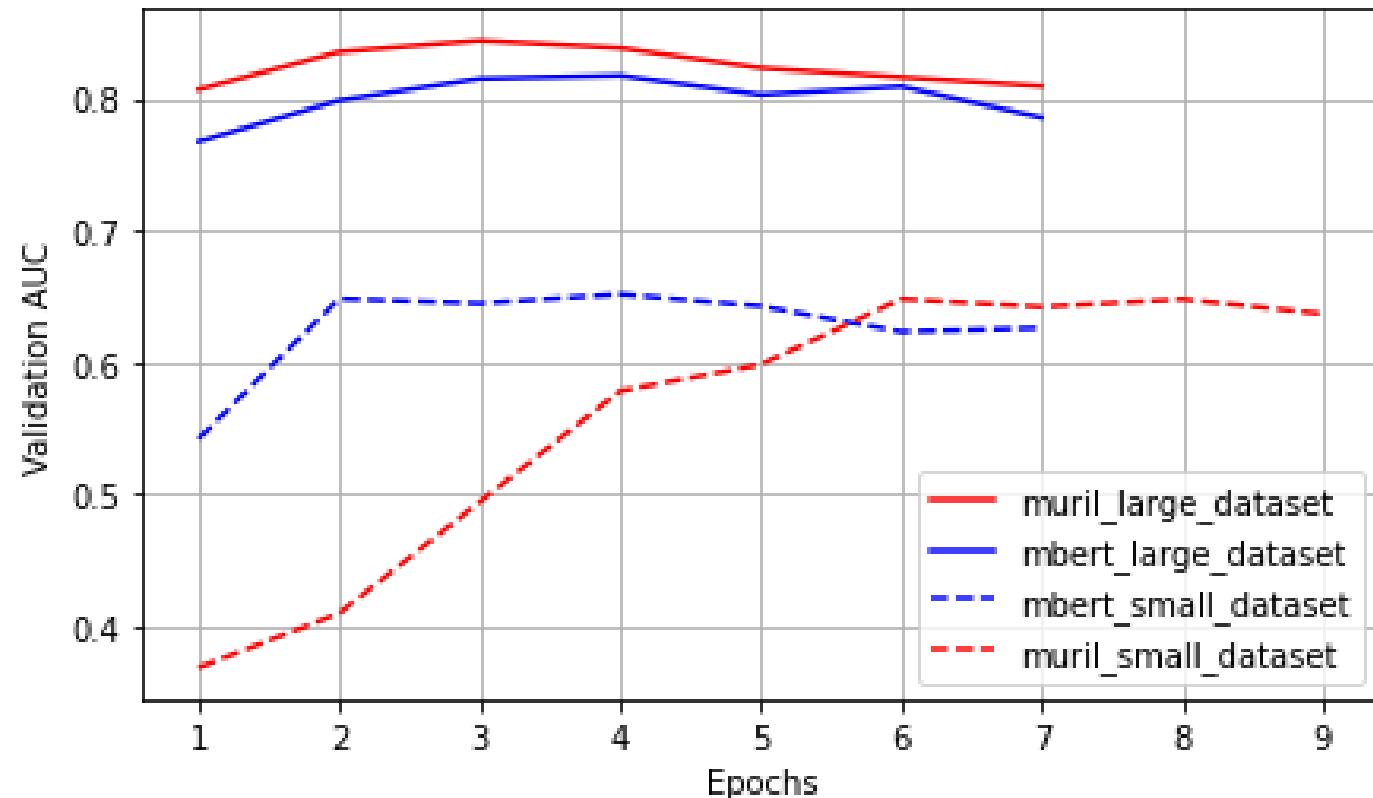


Metrics vs Kappa Score



mBERT vs MuRIL

- Language family-specific models better than generic models only when finetuning dataset size is of a certain minimum number.
- Small dataset: 8,089 tweets
- Large dataset: 12,241 tweets



Outline

- NAT-CTC Dataset
- Nepali Tweets Classification
- Results
- **Final Remarks**

Final Remarks

- Although our dataset is not large, it may be **valuable for transfer learning and semi-supervised learning**.
- Our dataset can help to make **multilingual datasets more inclusive**, and the **models trained on them more robust**.
- **Translating and transliterating to and from our dataset can help in augmentation** in various settings.



Thank You