SIGUL 2022. Marseille (FR), 24-25 June 2022

Question Answering Classification for Amharic Social Media Community Based Questions

Tadesse Destaw Belay[3], Seid Muhie Yimam[1]
Abinew Ali Ayele[1,2] and Chris Biemann[1]

[1]Universität Hamburg, [2] Bahir Dar University, [3] Wollo University

Outline

1. Introduction

Question classification, contributions

2. Data Collection and Processing

Telegram group, data processing, Latin to Ethiopic conversion

3. Model Discussion

Results, model comparison, error analysis

4. Conclusion

Contribution, dataset, future direction

01

Introduction

@AskAnythingEthiopia

Introduction

- Telegram public Channel
- More than 78k subscribers
- Exists since May 31, 2019
- Questions on various domains like politics, economics, health, education
- Questions posed in Amharic, English, Amharic-in-Latin

Contributions

- QAC dataset for Amharic
- Amharic-in-Latin to Ethiopic script transliteration
- Deep learning classification models
- Publicly release resources

02

Data Collection

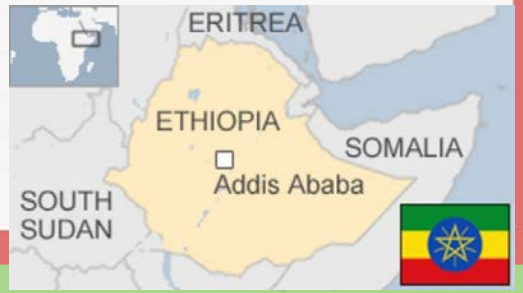
@AskAnythingEthiopia

Ask Anything Ethiopia

About Amharic

1. ሀ ha	12. ቸ che	23. የ ye
2. ለ le	13. ኀ kHa	24. ደ de
3. ሐ Ha	14. ነ ne	25. ጀ je
4. መ me	15. ኘ ñe	26. ገ ge
5. ሠ se	16. አ A	27. ጠ Te
6. ረ re	17. ከ ke	28. ፏ Che
7. ሰ se	18. ኸ He	29. ጰ P'e
8. ሸ she	19. ወ we	30. ጸ Ts'e
9. ቀ qe	20. ዐ ‘a	31. ፀ Tz'e
10. በ be	21. ዘ ze	32. ፈ fe
11. ተ te	22. ዠ zhe	33. ፐ p'e

- Second most Semitic language
- Written left to right in Ge'ez alphabets (Fidäl (ፊደል)/Ethiopic script)
- Syllable-based writing system where the consonants and vowels co-exist within each graphic symbol
  - Be(በ), Bu(ቡ).
- working language of the Federal Democratic Republic of Ethiopia



@AskAnythingEthiopia

Ask Anything Ethiopia

- Created by@JvHaile and @da_king

Rules

Question category, no fake news, no adds, community QA

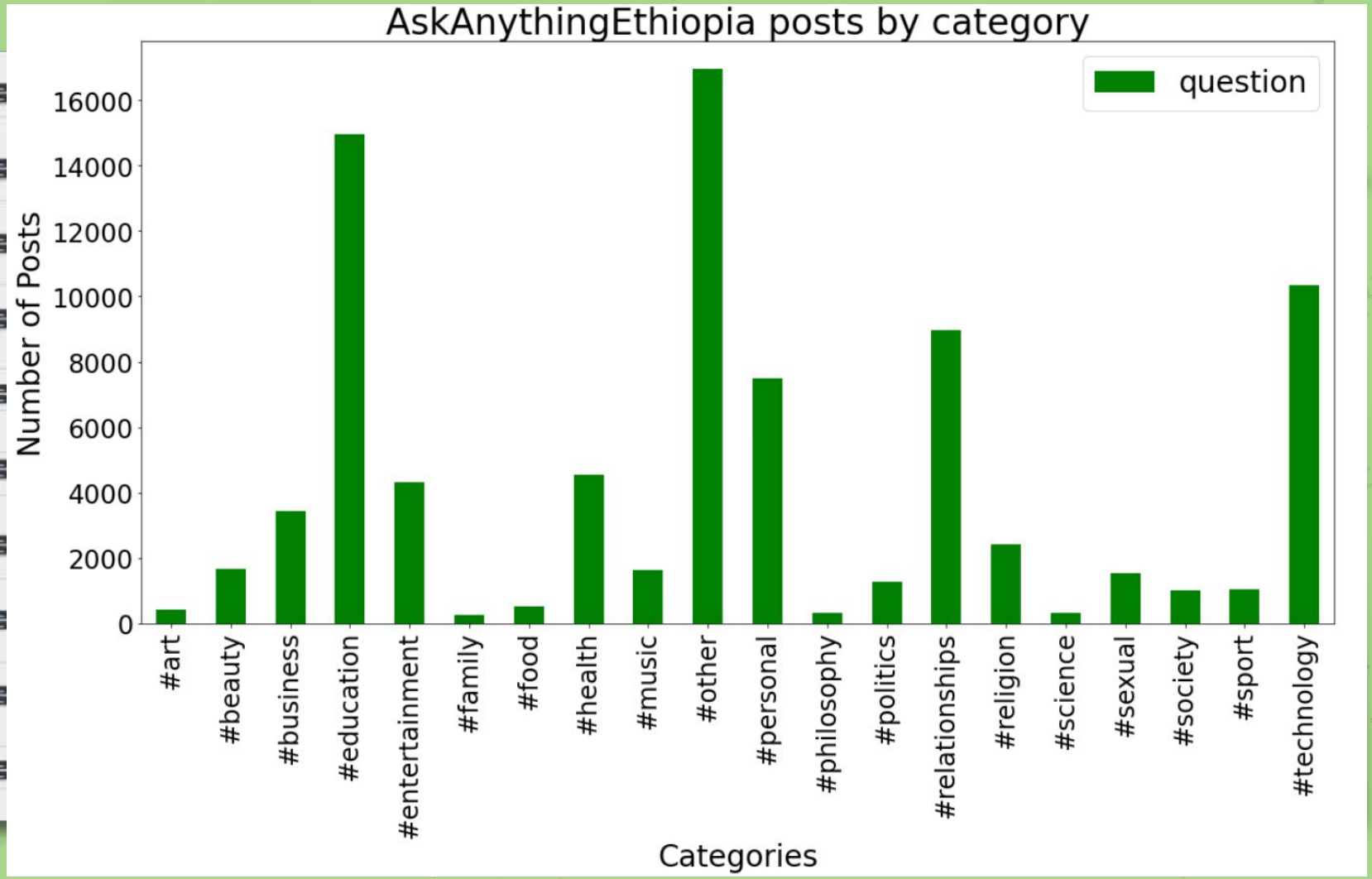
Question types

Politics, music, religion, …

Leaderboard

Reputation, privileged, award per month (500ETB)

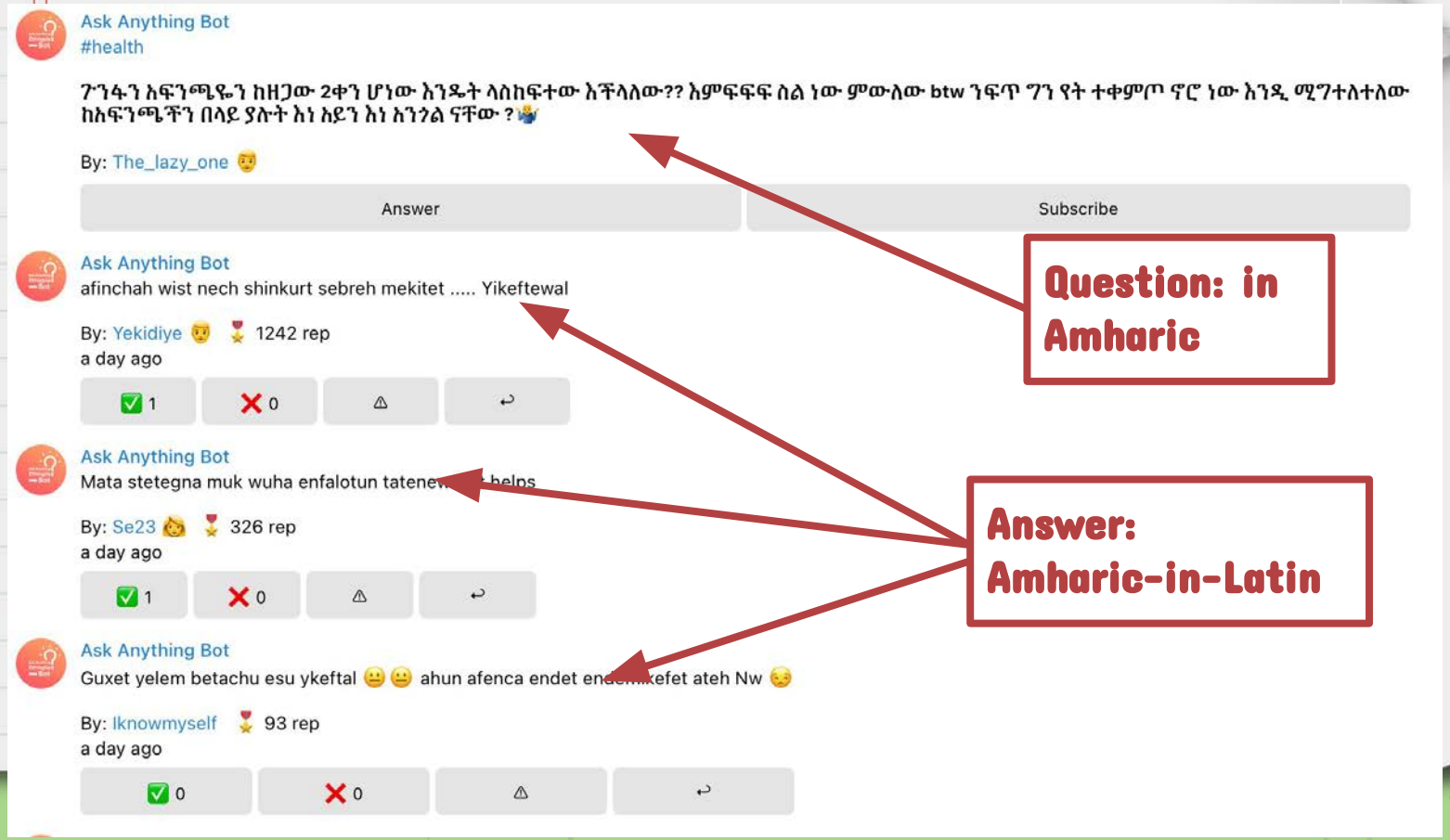

Ask Anything Bot
All Time Reputation Leaderboard
#1. Empire : 9593 rep
#2. DagiRelax : 8787 rep
#3. El_unico : 7858 rep
#4. Agenda : 6977 rep
#5. Marakiye : 6842 rep
#6. Bermel_mouth : 6742 rep
#7. PrOllfIC : 5977 rep
#8. Candor : 5294 rep
#9. Unknownn : 5069 rep
#10. Abadula : 5044 rep
You are at #217932 out of 296582 users


AskAnythingEthiopia posts by category
question
Number of Posts
0
2000
4000
6000
8000
10000
12000
14000
16000
#art
#beauty
#business
#education
#entertainment
#family
#food
#health
#music
#other
#personal
#philosophy
#politics
#relationships
#religion
#science
#sexual
#society
#sport
#technology
Categories

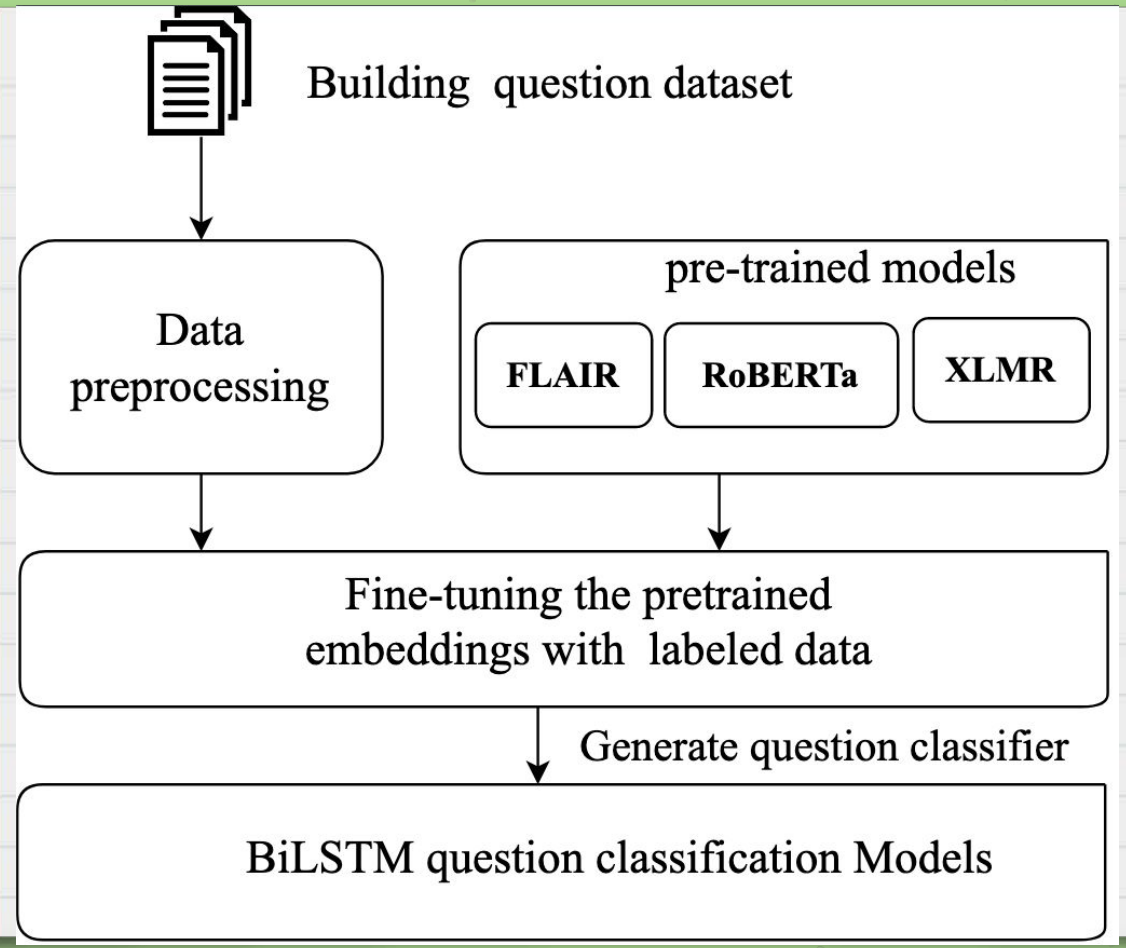
Posing a question

- User pose a question and should select the types
- Admin approves the question
- Other members start answering the question

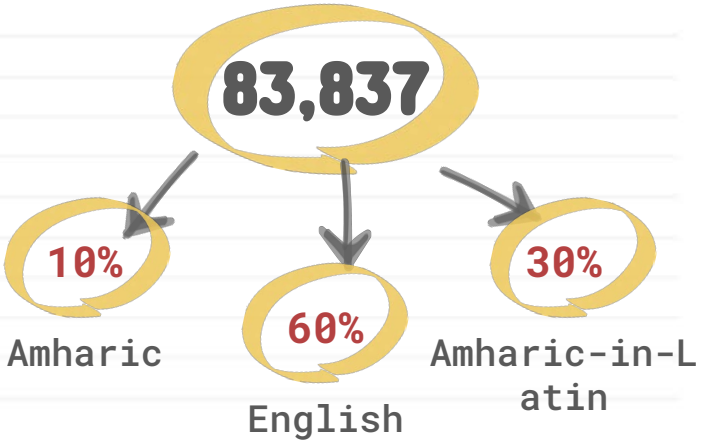
Question and Answer Example



System Pipeline



Data Collection



- Used Python Telethon library to get data from Telegram — 83,837 questions
- The Python Compact Language Detection library (CLD2) package — to detect languages - 7,967, 51,424, and 24,446 questions are posed in Amharic, English, and Amharic with a Latin script respectively.

Latin to Ethiopic Script Transliteration

Transliteration is a process of converting ASCII represented Amharic texts back to the canonical Amharic letter representations

"zare sint ken new?" ⇒ "ዛሬ ስንት ቀን ነው?"

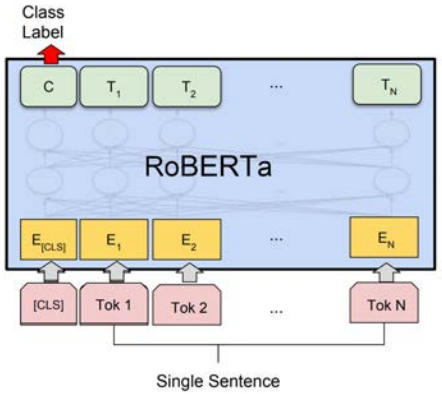
Parts of the Semantic Models for Amharic (https://github.com/uhh-lt/amharicmodels/)


```
From amseg.amharicTranslitrator import AmharicTranslitrator as  transliterator

transliterated = transliterator.transliterate('misa belah')
```

⇒ 'ሚሳ በላህ'

03



Model & Discussion

Classification Models

Three different contextual embedding approaches

- **AmRoBERTa**: Is a RoBERTa model that is trained for Amharic using 6.5m sentences (Yimam et al., 2021)
- **AmFLAIR**: FLAIR embedding using same data above
- **XLMR**: (Conneau et al., 2019)

- **AmRoBERTa** works better than others
- **XLMR** fails for almost all categories except Others

Amharic Questions		RoBERTa			AmFLAIR		
Q. Categories	No. of Q.	P	R	F1	P	R	F1
Education	1118	63.71	68.70	66.11	59.26	69.57	64.00
Personal	763	27.71	28.40	28.05	24.49	14.81	18.46
Relationships	684	71.88	74.19	73.02	60.47	83.87	70.27
Technology	681	71.15	52.86	60.66	58.57	58.57	58.57
Religion	305	70.59	68.57	69.57	73.97	77.14	75.52
Health	519	54.55	67.92	60.50	50.00	62.26	55.46
Business	363	34.78	47.06	40.00	33.33	32.35	32.84
Entertainment	305	14.29	16.67	15.38	30.77	22.22	26.81
Politics	269	67.86	76.00	71.70	57.58	76.00	65.52
Music	218	47.62	66.67	55.56	43.75	46.67	45.16
Society	194	22.22	21.05	21.62	00.00	00.00	00.00
Beauty	125	40.00	23.53	29.63	100.0	11.76	21.05
Sexual	108	100.0	42.86	60.00	100.0	28.57	44.44
Philosophy	102	33.33	44.44	38.10	00.00	00.00	00.00
Sport	93	70.00	46.67	56.00	100.0	26.67	42.11
Art	56	66.67	50.00	57.14	00.00	00.00	00.00
Food	53	33.33	25.00	28.57	00.00	00.00	00.00
Family	39	14.29	33.33	20.00	00.00	00.00	00.00
Science	24	20.00	100.0	33.33	00.00	00.00	00.00
Other	1518	30.71	33.75	32.49	31.75	41.88	36.12
Av. f1 (micro)				50.82			48.93
Av. f1 (macro)				46.07			32.77

Error Analysis: Transliteration mistake

Example 1

Original: Hi menjafekad lemawtat ke sent amet jemro new?

Transliterated: ሂ መንጃፈካድ ለማውታት ከ ሰንት አመት ጀምሮ ነው?

English: Hi, what is the minimum age to obtain a driving licence?

Error Analysis: Model was Correct!

Example 2

Amharic: **ሰላም ስለ ኤርትራ እንደ ሀገር መመስረት በደንብ ሚገልፅ መፅሃፍ ጠቁሙኝ እባካችሁ?**

Translation: Hi, Please tell me a book that clearly describes Eritrea as a nation

- Gold: education
- Pred: politics

Example 3

Amharic: **አልወደኩም በፈራሁት ላይ የምለውን መዝሙር ላኩልኝ እባካችሁ?**

Translation: Please send me a Mezmur (religious song) entitled as I did not fail on what I was scared of?

- Gold: music
- Pred: religion

Error Analysis: Model was wrong!

Example 4

Amharic: ያፈቀሩትን ሠው መርሣት ይቻላል ይባላል እንዴት መርሳት ይቻላል?

Translation: It is said that the person you love can be forgotten. How to forget?

- Gold: relationships
- Pred: technology

Example 5

Amharic:አሁን በዚህ ሰአት ምን እየተሰማቹ ነው?

Translation:what are you feeling right now?

- Gold: other
- Pred: politics

Cross-dataset Model Evaluation

Model	Test	P	R	F1
Merged	Amharic	47.61	39.65	42.11
Amharic	Trans.	25.71	20.37	21.44
Merged	Trans.	42.24	37.78	39.39
Trans.	Amharic	43.67	34.33	36.92

04

Conclusion & Future work

Conclusion

- Question classification dataset from community based social media platform
- The RoBERTa model trained on Amharic text works better than others
- Cross-dataset (Amharic, Transliterated, and merged) models perform less compared to in-dataset evaluation
- Resources publicly released

Future works

- End-to-end QA system
- Multilingual question classification (Amharic + English)
- Improving the transliteration system
  - using a dictionary
  - contextual embeddings for word correction
- Multi-modal system (images, sounds,and videos)

Thank you for your attention

Amharic Semantic Models

https://github.com/uhh-lt/amharicmodels/

SCAN ME

Semantic Models for Amharic

AMHARIC NLP
BENCHMARK RESOURCES

The semantic models resources are added to Lanfrica

amharic-corpus

Semantic models

Amharic Segmenter, toknizer, and translitrator

Announcements

🎉🎉🎉 The Amharic RoBERTa model is uploaded in Huggingface Amharic RoBERTa