Proceedings of SIGUL2022 @LREC2022, pages 137–145
Marseille, 24-25 June 2022

Question Answering Classification for Amharic Social Media Community Based Questions

Tadesse Destaw Belay[1], **Seid Muhie Yimam**[2], **Abinew Ali Ayele**[2, 3], **Chris Biemann**[2]
Wollo University, Dessie, Ethiopia[1],
Universität Hamburg, Hamburg, Germany[2],
Bahir Dar University, Bahir Dar, Ethiopia[3]
tadesseit@gmail.com, {seid.muhie.yimam, abinew.ali.ayele, christian.biemann}@uni-hamburg.de

Abstract

In this work, we build a Question Answering (QA) classification dataset from a social media platform, namely the Telegram public channel called @AskAnythingEthiopia. The channel has more than 78k subscribers and has exists since May 31, 2019. The platform allows asking questions that belong to various domains, like politics, economics, health, education, and so on. Since the questions are posed in a mixed-code, we apply different strategies to pre-process the dataset. Questions are posted in Amharic, English, or Amharic in Latin script. As part of the pre-processing tools, we build a Latin-to-Ethiopic-Script transliteration tool. We collect 8k Amharic and 24K Amharic but written in Latin script questions and develop deep learning-based questions answering classifiers that attain an F-score of 57.79 in 20 different question categories. The datasets and pre-processing scripts are open-sourced to facilitate further research on the Amharic community-based question answering.

Keywords: question answering, Latin transliteration, question classification, Amharic question answering, social media questions

1. Introduction

Question classification (QC) is growing in popularity as it has an important role in Question Answering (QA) systems, and Information Retrieval (IR) and it can be used in a wide range of other domains (Sangodiah et al., 2015). The main aim of question classification is to accurately assign labels to questions based on the expected answer type (Metzler and Croft, 2005). It plays an important role in finding or constructing accurate answers and therefore helps to improve the quality of automated question answering systems (Van-Tu and Anh-Cuong, 2016). To correctly answer a question, one needs to understand what the question asks for.

Moreover, question classification, which focuses on putting the questions into several semantic categories, can minimize constraints on the possible answers and suggest different processing strategies. For example, if the system understands the question "Who will win the Presidential election?" asks for a person name from a "politics" category, the search space of possible answers will be significantly reduced. It aims to solve answer generating issues by extracting the relevant features from the questions and by assigning them to the correct class category. More specifically, knowing the possible classes of the question before answering narrows down the number of possibilities a question answering system has to consider (May and Steinberg, 2004).

While there are some attempts in building question answering systems for Amharic (Yimam and Libsie, 2009; Taffa and Libsie, 2019; Abedissa, 2013), as far as we know, there are no publicly available datasets for question classification tasks. To address this gap, we have collected question answer datasets from a social media platform community question and answer channel. The @AskAnythingEthiopia[1] Telegram channel has been established in 2019, where users are allowed to ask questions of various categories such as science, education, religion, art, and so on. Figure 2 shows the distributions of questions per different question classes or categories. The community give answers for each question, which is governed by administrators of the channel.

The main contributions of this work are:

1. Introduce the first public question answering classification dataset for Amharic.
2. Implement a transliteration algorithm that converts questions written in Latin script to Amharic Ethiopic or Fidäl representation.
3. Build deep learning models to classify questions into pre-defined categories.
4. Investigate the quality of the different question categories that have been collected from

[1] https://t.me/askAnythingEthiopia

the social media platform.

5. Publicly releases the QA dataset along with the Amharic semantic models and resource repository (Yimam et al., 2021).

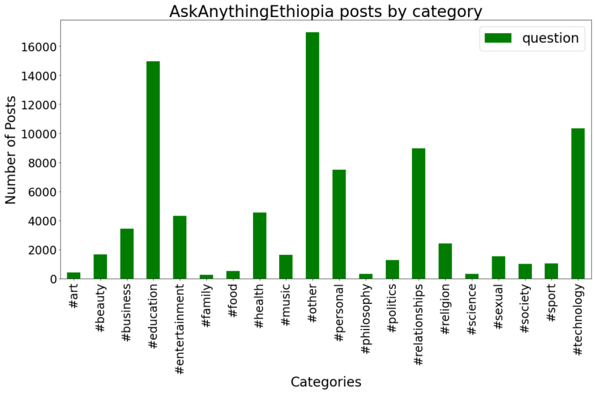


Figure 1: Distribution of questions per question categories.

In Section 2, basic information about the Amharicn language and the writing systems are discussed. In Section 3, we have presented the related works about question classification and some of the existing question answering systems for Amharic. While Section 4 and 5 discussed the data collection and pre-processing strategies, we have presented the Latin to Ethiopic/Fidäl transliteration processes in Section 6. In Section 7 and 8, deep learning question classification models and the results obtained are discussed. Finally, we have presented the main finding and future works in Section 9.

2. Amharic Language

Amharic (**አማርኛ**, amarəñña) is written from left to right in Ge'ez alphabets called Fidäl (**ፊደል**), also known as Ge'ez or Ethiopic script (Amha, 2009). Fidäl is a syllable-based writing system where the consonants and vowels co-exist within each graphic symbol. Amharic is the working language of the Federal Democratic Republic of Ethiopia and for many regional states in the country. It is the second old-most commonly spoken Semitic language after Arabic. Including the vowels, there are a total of 34 major letters each having up to seven major derivatives. Amharic uses a total of more than 300 characters.

3. Related Works

Many studies have addressed the question classification tasks, especially for high-resource languages like English. Among these, the work done by (Van-Tu and Anh-Cuong, 2016; May and Steinberg, 2004; Li and Roth, 2006; Li and Roth, 2002) proposed a method of using a feature selection algorithm to determine appropriate features corresponding to different question types. These proposed approaches are also used by the Text REtrieval Conference (TREC) shared task. The TREC dataset[2] is for question classification consisting of open-domain, fact-based questions divided into broad semantic categories. It has both a six-class called TREC-6, namely, Abbreviation, Description, Entities, Human Beings, Locations, or Numeric Values, and a fifty-class (TREC-50) version. Lei et al. (2018) proposed a novel CNN-based method for question classification in intelligent question answering using 5 different dataset types to test the performance of the proposed method. The work by Yang et al. (2018) built an attention-based LSTM to conduct Chinese questions classification. This work used Fudan University's question classification dataset, including 17,252 Chinese questions and classification results. Even though QC has been studied for various languages, it was barely studied for Amharic language and there is no benchmark dataset for question categorization. The work by Nega et al. (2016) presented Amharic question classification using machine learning (SVM) approaches. However, the dataset set used in this research consists of a very small dataset and is not publicly available, where a total of 180 questions are collected from the Agriculture domain.

Habtamu (2021) prepared an Amharic question dataset by labeling the sample questions into their respective classes and implemented an Amharic Question Classification (AQC) model using Convolutional Neural Network (CNN). The collected dataset was around 8,000 generic Amharic questions from different websites and labeled into 6 classes, similar to the question classes proposed by Li and Roth (2006). However, the dataset is still not available for further investigation. The work by Taffa and Libsie (2019) and Abedissa (2013) have developed Amharic non-factoid QA for biography, definition, and description questions. Yimam and Libsie (2009) developed an Amharic question answering system for factoid questions.

To the best of our knowledge, there are no publicly available question classification datasets that address the growing community-based question and answer platforms. We have collected the largest Amharic QC dataset to date.

4. Data Collection

One of the big challenges for low-resource languages such as Amharic is the unavailability of general-purpose datasets for various NLP tasks. For the question answering task, there is no publicly available benchmark dataset for Amharic. Some of the QA tasks, such as those by Yimam and Libsie (2009) and Nega et al. (2016) dealt

[2] http://l2r.cs.uiuc.edu/cogcomp/Data/QA/QC/

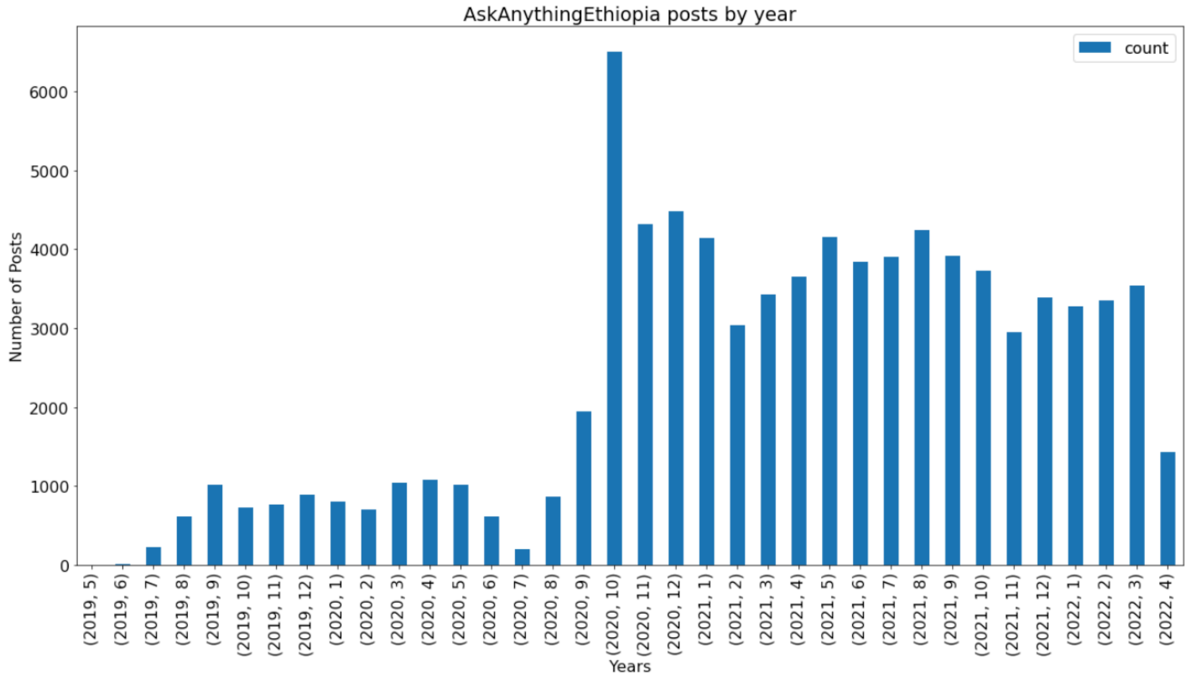
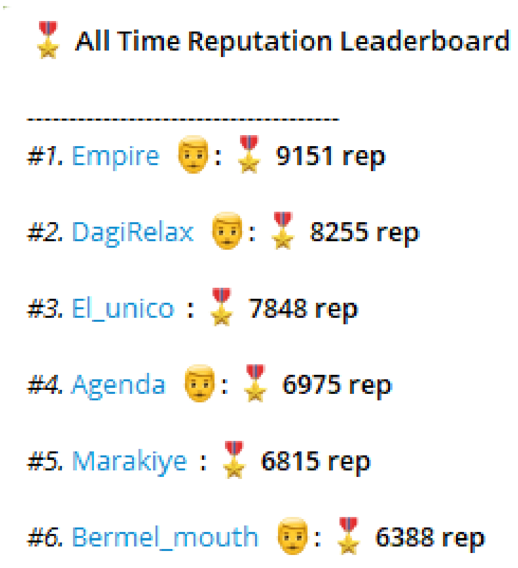


Figure 2: Distribution of questions over the last three years per each month (**year**, **month**).

with an end-to-end QA pipeline for factoid and domain-specific questions based on specific patterns. However, to build machine learning-based QA systems, manually annotated datasets are required. In this work, to build the QC datasets, we have exploited an existing social media platform community-based question and answer channel. Among several social media platforms, Telegram is one of the fastest-growing social networks platform in Ethiopia that has different features like bot services, personal chatting, and group calling/messaging. We have collected the Amharic question dataset from the public Telegram group channel called @AskAnythingEthiopia. The questions are freely available to the public who joined the group.

4.1. About @AskAnythingEthiopia

This Telegram group was created by @JvHaile and @da_king Telegram users. It was created for only questions that can not be answered with a simple Google search. Among the rules, 1) users are suggested to select the proper question category, 2) do not spread false information, 3) do not use it for announcements, and 4) don not ask questions that can be answered with a simple Google search. If users violate one of the rules, they will not be approved to ask further questions. Users that do not adhere to the rules will receive a warning, and if they continue breaking the rules, they will be banned from the channel permanently. It is the first of its kind in Ethiopia that serves only question answering in Amharic and/or English languages, which is a reward-based channel. Figure 3 shows the top 6 all-time leaders in reputation from the group.



🎖 All Time Reputation Leaderboard

#1. Empire 👨: 🎖 9151 rep

#2. DagiRelax 👨: 🎖 8255 rep

#3. El_unico : 🎖 7848 rep

#4. Agenda 👨: 🎖 6975 rep

#5. Marakiye : 🎖 6815 rep

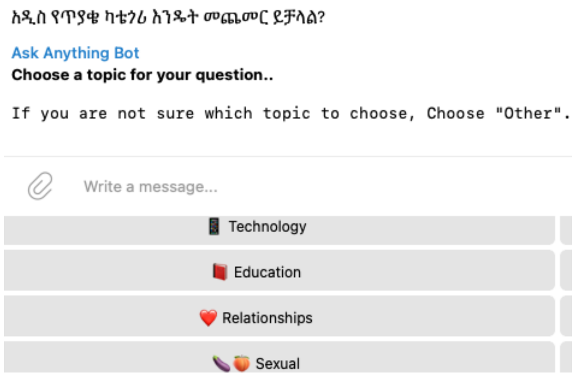
#6. Bermel_mouth 👨: 🎖 6388 rep

Figure 3: Top 6 all-time reputation leaders of the bot (accessed on 18 April 2021).

Reputation is the number of points that each user has obtained weekly, monthly, and all time. It is an indicator of how helpful their answers were as well as how often the answers were seen. The more reputation they have, the more privileges they have on the bot. For example, asking an unlimited amount of questions per day depends on the reputation. In addition to this, they will also be eligible to be rewarded with 500 Ethiopian Birr at the end of each month.

4.2. Posing a Question

The question is asked to a bot under the group called @ask_anything_ethiopia_bot. At the time of writing, this bot has 287,557 subscribers. Figure 4 shows the user interface displayed by the bot to facilitate asking a question and selecting the appropriate question categories. Once the user entered the */start* command, the bot is initiated and displays the list of options including **Ask a question**. If the question type does not fall in one of the existing 20 categories, users are forced to select the category "other". Once the questions are posed to the group, they will be displayed under the @AskAnythingEthiopia channel where users respond to the questions. Using the Python Telethon[3] library, we have extracted 83,851 questions with their categories. Figure 1 shows the distribution of questions per question class or category while Figure 2 shows the number of questions over the past years. As we can see from Figure 2, the number of questions asked in the channel increases over time.



አዲስ የጥያቄ ካቴጎሪ እንዴት መጨመር ይቻላል?

Ask Anything Bot
Choose a topic for your question..

```
If you are not sure which topic to choose, Choose "Other".
```

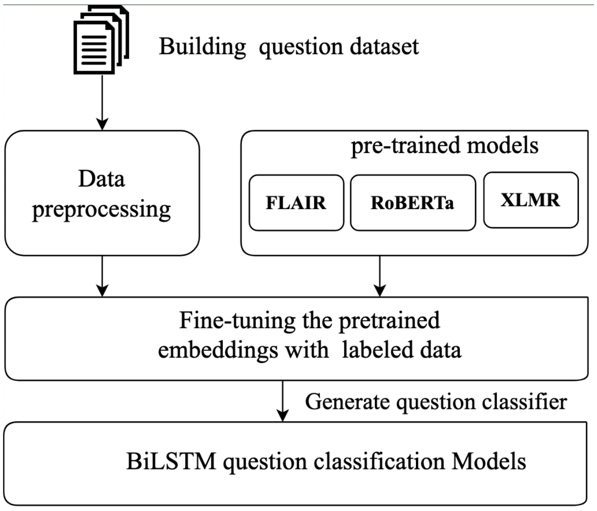
Figure 4: @ask_anything_ethiopia_bot Telegram bot user interface to ask questions.

5. Data Pre-processing

The platform allows asking questions both in English and Amharic. We have found that the questions are asked in different forms such as 1) all questions in Amharic, 2) all questions in English, 3) questions mixed in English and Amharic, or 4) Questions asked in Amharic language but written in Latin script.

The Python Compact Language Detection library (CLD2)[4] package is used to detect the script of the questions and we have found that 7,967, 51,424, and 24,446 questions are posed in Amharic, English, and Amharic with a Latin script respectively.

In this study, we have considered questions written in Amharic Fidäl or Latin scripts to build the machine learning models. In the future, the questions posed in English will be used to build a multilingual question classification model. For questions written in the Latin script, we have implemented an algorithm that tries to convert the text to its nearest possible Amharic Fidäl representation, as discussed in Section 6 below.



Building question dataset
Data preprocessing
pre-trained models
FLAIR
RoBERTa
XLMR
Fine-tuning the pretrained embeddings with labeled data
Generate question classifier
BiLSTM question classification Models

Figure 5: A general framework for the proposed Amharic question classification

6. Latin to Ethiopic Script Transliteration

Due to various reasons, users prefer to write Amharic text in Latin scripts. The following are some of the probable reasons to use the Latin script for Amharic text: 1) the mobile or computer keyboard does not support Ethiopic scripts, 2) writing in Latin script is faster than using the Ethiopic keyboard which usually requires multiple keystrokes for a single character representation, and 3) most of the emojis and special character representation are easier to type using the English keyboards. Our analysis shows users prefer to write using the Latin script as much as three-time (24,446 questions) compared to using the Ethiopic scripts (7,967 questions).

There is no word embedding or transformer-based language models for Amharic text written in Latin scripts. Hence, in this work, we have implemented the first Latin to Ethiopic transliteration algorithm and publicly release the script alongside the **amharicprocessor**[5] Amharic text segmentation, normalization, and romanization tool (Belay et al., 2021) which is one of the resources built along with the Amharic semantic models (Yimam et al., 2021)[6]. Transliteration is a process of converting ASCII represented Amharic texts back to the canonical Amharic letter representations (which

[3] https://github.com/LonamiWebs/Telethon

[4] https://pypi.org/project/pycld2/

[5] https://pypi.org/project/amseg/

[6] https://github.com/uhh-lt/amharicmodels/

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	39.71	33.75	36.49	31.75	41.88	36.12
Av. f1 (micro)				50.82			48.93
Av. f1 (macro)				46.07			32.77

Table 1: Amharic question distributions and classification model results using AmRoBERTa and AmFLAIR embeddings.

are known as Ethiopic or Fidäl scripts). For example, the phrase "zare sint ken new?" written in classical Latin script can be transliterated to its Ethiopic representation as "ዛሬ ስንት ቀን ነው?" (Translation: what is the date of today?).

To transliterate Latin-based Amharic texts to their Fidäl/Ethiopic based Amharic representation, we have constructed rules, that try to reproduce the Ethiopic representation with minimal errors, as a perfect reproduction is difficult. The rule is compiled with a list containing the ASCII combinations and the corresponding Amharic letters where the largest possible chunk are first transliterated before transliterating smaller units. For example, we first look for sh (ሽ) before attempting to transliterate s (ስ).

It should be noted that the transliteration effort is different from the standard International Phonetic Alphabet (IPA) representation (Tedla, 2015), as users generally ignored the IPA pronunciation of words in different accents.

Example **1** shows an Amharic question from our dataset posed in a Latin script; the 'Original' line is the original question, and the 'Transliterated' line is the question transliterated to its Fidäl script equivalent, while the 'English' line is the translation of the given question to English. The red colored text indicated errors introduced by the transliteration algorithm. Here, the first error is introduced as the word is written in English (Hi) while the remaining errors are introduced because the Amharic characters ቀ and ጠ have similar representation in the non-IPA Latin script with ከ and ተ, which are **ke** and **te** respectively.

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?

7. Classification Models

A great deal of current research works on question classification are based on deep learning approaches with contextual embeddings rather than statistical approaches. In this experiment, we have employed three different contextual embedding approaches, where two of them are from the Amharic Semantic resource repository (Yimam et al., 2021) while the third one is from a publicly available embedding model from HuggingFace[7].

1. XLMR: Unsupervised Cross-lingual Representation Learning at Scale (XLMR) is a generic cross-lingual sentence encoder that is trained on 2.5 TB of newly-created clean CommonCrawl data in 100 languages including

[7] https://huggingface.co/xlm-roberta-base

Amharic (Conneau et al., 2019). Among this, 68m tokes are for Amharic.

2. AmRoBERTa: Is a RoBERTa model (Liu et al., 2019), that is trained for Amharic using a 6.5m sentences crawled from different sources (Yimam et al., 2021).

3. AmFLAIR: is based on FLAIR, a framework designed to facilitate experimentation with different embedding types, as well as training and distributing sequence labeling and text classification models (Akbik et al., 2018). This is a new FLAIR embedding model that was trained from scratch using a 6.5m Amharic corpus (Yimam et al., 2021).

AmRoBERTa and AmFLAIR embedding models are publicly available on GitHub[8] with the different benchmark datasets and NLP models.

A general framework using the deep learning method for our question classification is shown in Figure 5. As shown in the diagram, first, we need to build a question classification training dataset scraped from @AskAnythingEthiopia Telegram public channel. For all experiments, the data are further split into training, development, and test instances using an 80:10:10 split.

We have fine-tuned the pre-trained transformer/contextual pre-trained language models using the question classification datasets using a BiLSTM-based text classification model from FLAIR. The Text classification architecture is composed of respective embedding layers as an input layer with the sequence of 4 dense layers and an output layer. The training parameters for the architecture constitute a learning_rate of 0.5e5, mini_batch_size of 4, and max_epochs of 10. The models are trained on a 'Quadro RTX 6000' GPU server. While the Amharic dataset training took about 3 hours, the transliterated and merged (transliterated and Amharic) training took about half a day. We did not use the English dataset for model training.

The experimental results for the three different datasets (Amharic, Transliterated, and Merged) using the models fine-tuned on the two pre-trained embeddings (AmRoBERTa and AmFLAIR) are shown in Table 1, 2, and 3 respectively. Since the finetuned model based on XLMR could not produce meaningful results (it miss classify almost all of the cl assess, except the "others" class), we have excluded the results from the Tables. The cross-evaluation of the different models are shown in Table 4.

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

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

8. Discussion

In this section, we will discuss the results of the Amharic question classification experiments we have presented in Section 7. We have used the F1-score (F1), Precision (P), and Recall (R) for the comparison of the models' performances for each question class. For the overall performances of the models, we have reported the average micro F1-scores as it shows us the overall performance. For completeness, we have reported the average macro F1-scores, but the scores will not be concrete as the classes are not balanced. The models fine-tuned on the AmRoBERTA pre-trained model have achieved an F1 score of 57.29 while those on AmFLAIR have achieved an F1 score of 54.20. Models fine-tuned from the multi-lingual XLMR embedding could not able to predict the question classes at all, except for the "Others" class with an F1 score of less than 20%. Hence, we have excluded the results from all tables.

When we see the results at the class label, questions under **Politics** and **Religion** classes are relatively accurately predicted. The class on **Entertainment** is the worst classified by the models. The class under **Other** has more questions than the other class but still, the model wrongly predicts most of the questions. One possible reason could be that the questions under **Other** are not seman-

[8] https://github.com/uhh-lt/amharicmodels

Transliterated Questions		RoBERTa			AmFLAIR		
Q. Categories	No. of Q.	P	R	F1	P	R	F1
Education	4542	74.09	79.36	76.63	64.65	78.44	70.88
Personal	2127	23.78	22.00	22.86	30.97	17.50	22.36
Relationships	3007	76.70	81.72	79.13	66.37	77.59	71.54
Technology	2703	70.55	71.03	70.79	65.46	68.62	67.00
Religion	933	76.47	66.67	71.23	56.96	57.69	57.52
Health	1427	65.71	62.59	64.11	50.98	53.06	52.00
Business	861	42.03	31.52	36.02	28.95	11.96	16.92
Entertainment	880	36.59	32.61	34.48	33.33	14.13	19.85
Politics	296	47.62	41.67	44.44	72.73	33.33	45.71
Music	761	61.97	69.84	65.67	53.32	69.84	61.11
Society	254	06.67	05.26	05.88	00.00	00.00	00.00
Beauty	571	43.48	33.33	37.74	41.67	08.33	13.89
Sexual	325	50.00	41.38	45.28	00.00	00.00	00.00
Philosophy	50	11.11	100.0	20.00	00.00	00.00	00.00
Sport	300	34.78	23.53	28.07	42.86	08.82	14.63
Art	116	50.00	36.36	42.11	00.00	00.00	00.00
Food	144	14.29	20.00	16.67	00.00	00.00	00.00
Family	81	14.29	16.67	15.38	00.00	00.00	00.00
Science	41	33.33	16.67	22.22	00.00	00.00	00.00
Other	4542	41.57	44.71	43.08	38.72	53.78	45.03
Av. f1 (micro)				57.29			53.47
Av. f1 (macro)				42.09			27.91

Table 2: Transliterated classification model results using AmRoBERTa and AmFLAIR embeddings

Mixed Questions		RoBERTa			AmFLAIR		
Q. Categories	No. of Q.	P	R	F1	P	R	F1
Education	5660	70.16	78.95	74.30	63.88	80.58	71.27
Personal	2890	23.73	26.69	25.13	30.30	17.79	22.42
Relationships	3691	75.96	78.98	77.44	66.59	78.12	71.90
Technology	3384	69.49	68.33	68.91	66.04	68.61	67.30
Religion	1238	72.39	65.54	68.79	66.67	68.92	67.77
Health	1946	60.39	62.50	61.43	54.75	60.50	57.48
Business	1224	41.75	34.13	37.55	40.48	26.98	32.38
Entertainment	1185	44.29	28.18	34.44	32.73	16.36	21.82
Politics	565	56.00	57.14	56.57	59.57	57.14	58.33
Music	979	66.18	57.69	61.64	54.95	64.10	59.17
Society	448	07.14	07.89	07.50	00.00	00.00	00.00
Beauty	696	41.18	27.27	32.81	44.44	15.58	23.08
Sexual	433	43.48	27.78	33.90	50.00	11.11	18.18
Philosophy	152	18.75	30.00	23.08	00.00	00.00	00.00
Sport	393	53.85	28.57	37.33	53.33	16.33	25.00
Art	172	42.11	34.78	38.10	00.00	00.00	00.00
Food	197	44.44	28.57	34.78	00.00	00.00	00.00
Family	120	00.00	00.00	00.00	00.00	00.00	00.00
Science	65	16.67	14.29	15.38	00.00	00.00	00.00
Other	6060	39.47	40.93	40.19	39.12	49.92	43.86
Av. f1 (micro)				54.77			54.20
Av. f1 (macro)				41.46			32.00

Table 3: The mixed of Amharic and transliterated question and classification model results using AmRoBERTa and AmFLAIR embeddings.

tically similar enough to each other, and hence, we suggest that the platform should allow the creation of new question categories by the users.

We have made some error analyses to explore the strength and weaknesses of the model as well as to see if there are issues in the datasets. As it can be seen from Examples **2** and **3**, the model predicts the questions correctly while the samples were wrongly annotated. Some possible explanations for this wrong annotation of such samples could be either the users did not understand the question classification task (Example **2**) or the question itself is ambiguous (Example **3** has the word 'music' but it specifically refers to religious songs).

When we analyzed the model predictions, the most miss-classified classes are from the "Other" class. From Example **4**, we can see the model wrongly classified the question as "Technology", even though there are no contexts provided regarding technology. Similarly, Example **5** is predicted as "Politics" even though the question does not have a clear connection to politics.

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

Table 4: The cross model evaluations results. "Trans." stands for Transliterated questions while 'Merged' stands for the merged questions (Amharic and transliterated). The hypothesis tested here is evaluating different models using different test sets.

Moreover, we have conducted a cross-model evaluation, mainly to verify the performance of the transliterated models. The results based on the pre-trained AmRoBERTa pre-trained embeddings are presented in Table 4. The results indicated that the Amharic model fails to properly classify transliterated texts while the transliterated model works better for Amharic test sets. Mixing the dataset increases the performance, but it is still very far from the performances of the models on the same dataset instances.

9. Conclusion and Future Works

In this paper, we presented the first work on the Amharic question classification task. Similar to the Reddit social news website, the @AskAnythingEthiopia Telegram public channel is established in 2019, which attracted as many as 78k subscribers to ask a question. The community asked any questions that could cover a wide range of question categories such as "Politics", "Music", "Technology", "Religion" and so on.

As the questions are manually tagged, and users are enforced to choose a category by the platform, it is a gold-standard dataset for question answering classification tasks. In this paper, we focused only on the question classification task. Since questions are asked both in English and Amharic, we apply language detection to consider questions only posed in Amharic. As most of the online community uses the Latin script to write Amharic questions, we also developed a Latin to Ethiopic transliteration algorithm. Using the cleaned dataset, we built deep learning-based question classification models using a pre-trained transformer and contextual embeddings. The question classification models performed at 57.79% F1 score on a total of 20 question categories, which is quite a promising result. The resources such as question classification datasets for Amharic, the models, transliteration and Pre-processing tools are released in our GitHub repository[9]. We anticipate that this dataset can be used and extended for several use-cases such as 1) extracting the answers and implementing an end-to-end QA system, 2) building multilingual question classification (Amharic + English) systems, 3) improving the transliteration system using a dictionary and contextual embeddings for word correction, 4) extracting the associated multi-modal data (images, sounds, and videos) to build a multi-modal QC and QA systems.

10. Bibliographical References

Abedissa, T. (2013). Amharic question answering for definitional, biographical and description questions. *Unpublished Master's Thesis, Computer Science Department, Addis Ababa University, Addis Ababa, Ethiopia.*

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649, New Mexico, USA.

Amha, A. (2009). On loans and additions to the Fidäl (Ethiopic) writing system. In *The Idea of Writing*, pages 179–196. Brill.

Belay, T. D., Ayele, A. A., Gelaye, G., Yimam, S. M., and Biemann, C. (2021). Impacts of homophone normalization on semantic models for Amharic. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 101–106, Bahir Dar, Ethiopia. IEEE.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv 2019 preprint arXiv:1911.02116*, page 8440–8451.

Habtamu, S. (2021). *Amharic Question Classification System Using Deep Learning Approach.*

[9] https://github.com/uhh-lt/amharicmodels: The dataset are released under a permissive license

Unpublished master thesis, Addis Ababa University.

Lei, T., Shi, Z., Liu, D., Yang, L., and Zhu, F. (2018). A novel cnn-based method for question classification in intelligent question answering. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–6, Sanya China.

Li, X. and Roth, D. (2002). Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, page 1–7, Taipei, Taiwan.

Li, X. and Roth, D. (2006). Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

May, R. and Steinberg, A. (2004). Al, building a question classifier for a TREC-style question answering system. *AL: The Stanford Natural Language Processing Group, Final Projects*.

Metzler, D. and Croft, W. B. (2005). Analysis of statistical question classification for fact-based questions. *Information Retrieval*, 8(3):481–504.

Nega, A., Chekol, W., and Kumlachew, A. (2016). Question classification in amharic question answering system: Machine learning approach. *International Journal of Advanced Studies in Computers, Science and Engineering*, 5(10):14–21.

Sangodiah, A., Muniandy, M., and Heng, L. E. (2015). Question classification using statistical approach: A complete review. *Journal of Theoretical & Applied Information Technology*, 71(3):386–395.

Taffa, T. A. and Libsie, M. (2019). Amharic question answering for biography, definition, and description questions. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 110–113, Florence, Italy. Association for Computational Linguistics.

Tedla, T. (2015). amLite: Amharic Transliteration Using Key Map Dictionary. *arXiv preprint arXiv:1509.04811*.

Van-Tu, N. and Anh-Cuong, L. (2016). Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, 9(17):1–8.

Yang, Y., Liu, J., and Liaozheng, Y. (2018). Chinese question classification based on deep learning. In *Advanced Multimedia and Ubiquitous Engineering*, pages 315–320. Springer.

Yimam, S. M. and Libsie, M. (2009). TETEYEQ: Amharic question answering for factoid questions. *IE-IR-LRL*, 3(4):17–25.

Yimam, S. M., Ayele, A. A., Venkatesh, G., Gashaw, I., and Biemann, C. (2021). Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).