Question Answering Classification for Amharic Social Media Community Based Questions

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

[1] College of Informatics, Wollo University, Kombolcha, Ethiopia
[2]Universität Hamburg, Hamburg, Germany
[3]Faculty of Computing, Bahir Dar University, Bahir dar, Ethiopia
tadesseit@gmail.com, {firtname.middlename.lastname}@uni-hamburg.de

Abstract

In this work, we build a Question classification (QC) dataset from a social media platform, namely the Telegram public channel called @AskAnythingEthiopia. The platform allows asking questions that belong to various domains, like Politics, Music, Technology, Religion and so on. Questions are posted in Amharic, English, or Amharic in Latin script. Since the questions are posed in a mixed-code, we apply different strategies to pre-process the dataset. 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 question 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.

1 Introduction

Question classification (QC) is growing in popularity as it has an important role in Question Answering (QA) systems and Information Retrieval (IR) (Sangodiah et al., 2015). The main aim of QC is to accurately assign labels to questions based on the expected answer type, improve the quality of automated QA systems (Metzler and Croft, 2005; Van-Tu and Anh-Cuong, 2016). While there are some attempts in building question answering systems for Amharic[1] (Yimam and Libsie, 2009; Taffa and Libsie, 2019; Abedissa, 2013), there are no publicly available datasets for question classification tasks. To address this gap, we have collected question datasets from a social media platform, @AskAnythingEthiopia Telegram question and answer channel.

[1]Amharic is the official language of the Federal Democratic Republic of Ethiopia (FDRE) and for many regional states in the country. It is written from left to right in Ge'ez alphabets called Fidäl.

The main contributions of this work are: 1) introduce the first public QC 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 predefined categories, and 4) investigate the quality of the different question categories that have been collected from the social media platform.

2 Methodology

Related Works: Many studies have addressed the QC tasks, especially for high-resource languages like English. Among these, the works by Van-Tu and Anh-Cuong (2016); May and Steinberg (2004); Li and Roth (2006, 2002); Lei et al. (2018) proposed methods of different feature selection algorithms to determine appropriate features corresponding to different question types. The TREC dataset 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 and a fifty-class (TREC-50) version. The work by Yang et al. (2018) built an attention-based LSTM to conduct Chinese questions classification. 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 works by Nega et al. (2016); Habtamu (2021); Taffa and Libsie (2019); Abedissa (2013); Yimam and Libsie (2009) presented Amharic question classification using different approaches. However, the dataset used is very small and is not publicly available.

Data Collection: In this work, to build the QC datasets, we have exploited an existing social media platform community-based question and answer channel. We have collected the Amharic question dataset from the public Telegram group channel called @AskAnythingEthiopia. Using the Python Telethon library, we have extracted 83851 ques-

tions with their categories. Figure 1 shows the distribution of questions per question class or category.

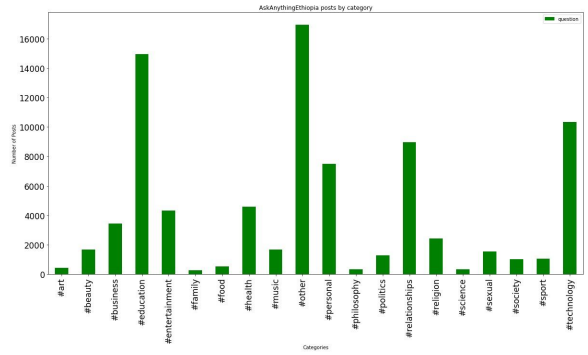

Figure 1: Distribution of questions per question categories.

The @AskAnythingEthiopia: This Telegram group has been established in 2019. It was created for only questions that can not be answered with a simple Google search and governed by rules. 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) do not ask questions that can be answered with a simple Google search. 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, the user more involved in the question and answering will rewarded with 500 Ethiopian Birr per month.

Data Pre-processing: The Python Compact Language Detection library (CLD2) package is used to detect the script of the questions and we have found that 7967, 51424, and 24446 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. 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.

Latin to Ethiopic Script Transliteration: Transliteration is a process of converting ASCII represented Amharic texts back to the canonical Amharic letter representations. 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.

Classification Models: In this experiment, we have employed three different contextual embedding approaches. These are: 1) Unsupervised Cross-lingual Representation Learning at Scale (XLMR), which is a generic cross-lingual sentence encoder that is trained on 2.5 TB of newly-created clean CommonCrawl data in 100 languages including Amharic (Conneau et al., 2019), 2) AmRoBERTa, a RoBERTa model (Liu et al., 2019), which is trained for Amharic using a 6.5m sentences crawled from different sources (Yimam et al., 2021), and 3) AmFLAIR (Yimam et al., 2021), is a FLAIR (Akbik et al., 2018) model that is trained for Amharic. We have fine-tuned the pre-trained transformer/contextual pre-trained language models using our QC datasets and trained a BiLSTM-based text classification model from FLAIR.

3 Experimental Setup and Results

For all experiments, the data are further split into train, development, and test instances using an 80:10:10 split. The training parameters for the model architecture constitute a learning_rate of 0.5e5, mini_batch_size of 4, and epochs of 10. The models are trained on a 'Quadro RTX 6000' GPU server. As show in Table 1, the classifiers trained using 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 trained using the multi-lingual XLMR embedding could not able to predict the question classes at all. When we see the results at the class label, questions under **Politics** and **Religion** classes are relatively accurately predicted. The class under **Other** has more questions but the model wrongly predicts most of the questions for this category.

Table 1: Experimental results (F1-score)

Question types	RoBERTa	AmFLAIR	XLMR
Amharic	50.82	48.93	1.68
Transliterated	**57.29**	53.47	1.65
Mixed	54.77	54.20	1.65

4 Conclusion and Future Works

In this paper, we presented the first work on the Amharic question classification (QC) task, where the data are collected from Telegram group called @AskAnythingEthiopia. The community asked any questions that could cover 20 categories. 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 QC models using a pre-trained transformer and contextual embeddings. The QC models from AmRoBERTa pre-trained embedding performed at 57.79% F1-score, which is quite a promising result. The resources such as QC datasets for Amharic, the models, transliteration and Pre-processing tools are available publicly in GitHub repository[2]. We anticipate that this dataset can be extended for several use-cases to explore Ethiopic NLP tasks[3] 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.

References

Tilahun Abedissa. 2013. Amharic question answering for definitional, biographical and description questions. *Unpublished Master's Thesis, Computer Science Department, Addis Ababa University, Addis Ababa, Ethiopia*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv 2019 preprint arXiv:1911.02116*, page 8440–8451.

Saron Habtamu. 2021. *Amharic Question Classification System Using Deep Learning Approach*. Unpublished master thesis, Addis Ababa University.

Tao Lei, Zhizhong Shi, Duoxing Liu, Lei Yang, and Feng Zhu. 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.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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

Adane Nega, Workneh Chekol, and Alemu Kumlachew. 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.

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

Tilahun Abedissa Taffa and Mulugeta Libsie. 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.

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

Yihe Yang, Jin Liu, and Yunlu Liaozheng. 2018. Chinese question classification based on deep learning. In *Advanced Multimedia and Ubiquitous Engineering*, pages 315–320. Springer.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).

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

[2] https://github.com/uhh-lt/ethiopicmodels

[3] https://github.com/EthioNLP