Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Language Technology

# AM-DETOX: Analyzing Amharic Text Detoxification Using Pre-trained Large Language Models

Abinew Ali Ayele[1,2], Seid Muhie Yimam[1]

[1]Language Technology Group, Department of Informatics, Universität Hamburg, Hamburg, Germany
[2] Bahir Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia

{abinew.ali.ayele, seid.muhie.yimam}@uni-hamburg.de

## Introduction

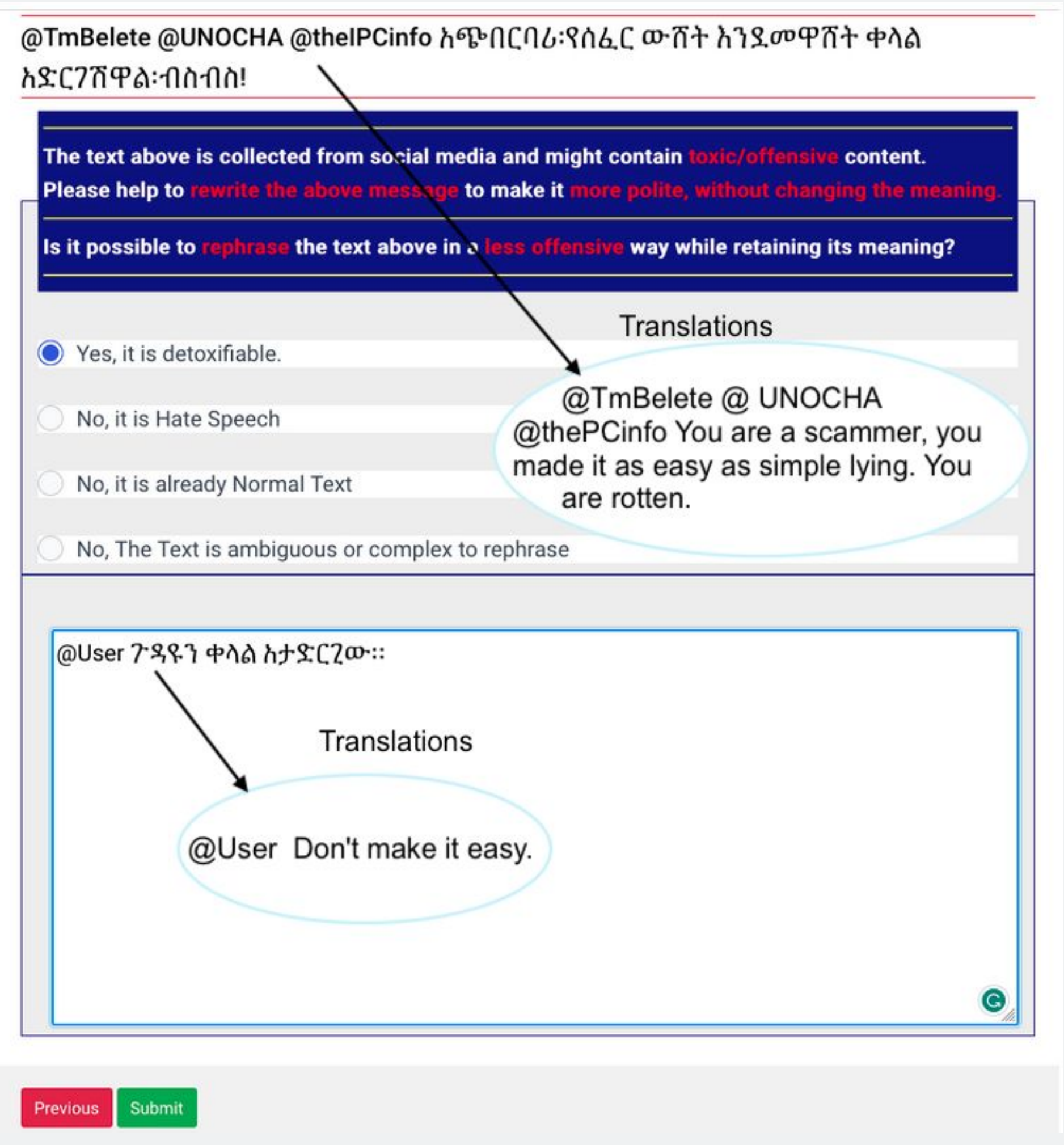
- Social media makes the spread of toxic content easier.
- Various mitigation strategies are employed.
- Most studies in low-resource languages focus on detection and classification tasks, fail to employ text detoxification
- Text detoxification or rewriting messages allows for the removal of toxicity from messages.
- Detoxification can protect vulnerable groups like children.
- LLMs were unable to detoxify toxic messages in low-resource languages, mainly "hallucinate".

## Research Questions

- **RQ1**: How effective are LLMs in detoxifying toxic content in low-resource languages such as Amharic?
- **RQ2**: The specific challenges in annotating toxic content and applying text detoxification techniques for the Amharic language?
- **RQ3**: How can the challenges be addressed?

## Data Collection

- **Source**: offensive labeled tweets from Ayele et al. (2022, 2023). https://github.com/uhh-lt/AmharicHateSpeech
- 3,120 tweets from both datasets
- Re-annotated for detoxification task



@TmBelete @UNOCHA @theIPCinfo አጭበርባሪ፡የሰፈር ውሸት እንዷመዋሸት ቀላል አድርገሽዋል፡ብስብስ!

The text above is collected from social media and might contain toxic/offensive content.
Please help to rewrite the above message to make it more polite, without changing the meaning.

Is it possible to rephrase the text above in a less offensive way while retaining its meaning?

● Yes, it is detoxifiable.
○ No, it is Hate Speech
○ No, it is already Normal Text
○ No, The Text is ambiguous or complex to rephrase

Translations
@TmBelete @ UNOCHA @thePCinfo You are a scammer, you made it as easy as simple lying. You are rotten.

@User ጉዳዩን ቀላል አታድርገው።

Translations
@User Don't make it easy.

Fig. 1: Annotation GUI

## Annotation

- Customized POTATO: POrtable Text Annotation Tool https://github.com/davidjurgens/potato
- Pilot annotation by 3 experts: **125** tweets
- Main annotation: **2,995** tweets
- Total annotations:
  - 1,452: detoxifiable
  - 1,543: non-detoxifiable

| | Toxic input |
|---|---|
| 1 | አንተ ቆሻሻ በዚህ ወቅት አይንህን ማየት አልፈልግም |
| | You garbage, I don't want to see your eyes at this time. |
| 2 | 2. እንዳንተ አይነት መጤ ዲቃላ ምንም አይመለከተውም በክት |
| | A bastard immigrant like you doesn't concern it, you carcass. |
| 3 | በአሳምነው በግፍ መገደል ብዙ ሆዳም ዋጋ መክፈሉ አይቀሬ ነው |
| | Many glutton people will pay price for the brutal murder of Asaminew. |
| 4 | 4. እናንተ አይደላችሁም እንዴ እዚህ የደረሳችሁን ። አረ እፈር የሆንክ እስስት |
| | Is that not you who got us here, shame on you, you are Chameleon. |

| | Detoxified annotation output |
|---|---|
| 1 | አንተ ጥሩ ሰው አይደለህም በዚህ ወቅት አንተን ማየት አልፈልግም |
| | You are not good a person, I don't want to see you at this time. |
| 2 | ይኸ አንተን አይመለከትም |
| | This does not concern you. |
| 3 | በአሳምነው መገደል ብዙ ሰው መጎዳቱ አይቀርም |
| | Many people may be affected for the murder of Asaminew |
| 4 | እናንተ ናችሁ እዚህ ያደረሳችሁን ፡ በዚህ ልትኮሩ አይገባም |
| | Is that not you who got us here, You should not be proud of this. |

Fig. 2: Text detoxification/ re-writing examples

## Experimental Results

- Am-ReBERTa achieved better performance on classification tasks

| Classifier | precision | recall | F1-score |
|---|---|---|---|
| Am-RoBERTa | **69.57** | **73.37** | **70.01** |
| Afro-xlmr-large | 69.07 | 73.53 | 67.75 |
| Bert-medium-amharic | 66.13 | 69.90 | 64.62 |

Table 1: Classification results in identifying detoxifiable texts from non-detoxifiable ones

- GPT4 and Shapely employed for identifying toxic words in text, where GPT4.0 mini outperformed.

| Toxic word detection performance | | Detoxification performance | |
|---|---|---|---|
| GPT-4o mini | SHAP | chrF2 | Sacrebleu |
| 76.33% | 54.83% | 11.5 | 0.5 |

Table 2: Toxic words detection (Explainability) and the detoxification capabilities (Detoxifiability) of SHAP and GPT-4 models.

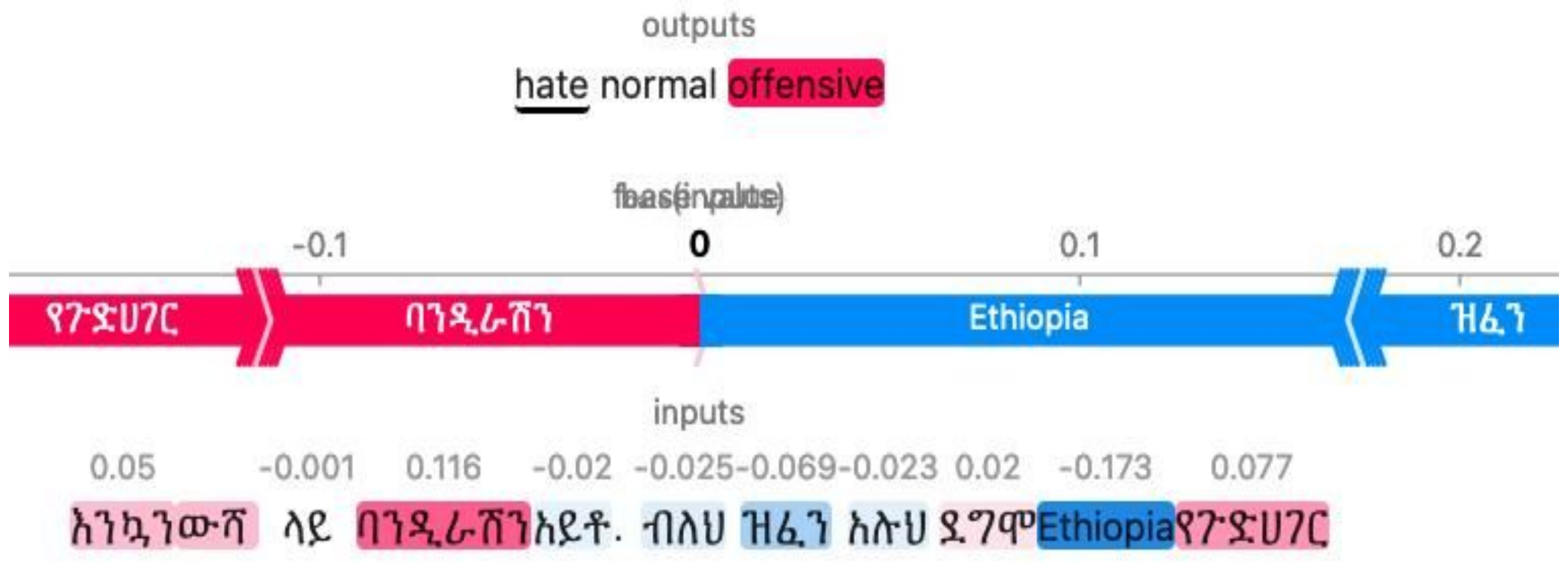


Fig.3: Output examples for explainable toxicity detection (SHAP)

- **GPT-4o-Mini** detoxification outputs significantly deviate from human expert annotations due to hallucinations.

| | Iput sentence | gpt4_detoxified | human_detoxified | Remark |
|---|---|---|---|---|
| 1 | ፈሳም (ፈሪ) ሰው አይመቸኝም ለሀገርም አይጠቅምም | ሁሉም ሰው ለሀገር ጥሩ ነው የሚሆን | ፈሪ የሆነ ሰው አልወድም ለሀገርም ይጠቅማል ብዬ አላስብም | The model halucinats & Expert do not detoxify well. |
| | A cowardly person is not comfortable, and it is not good for the country. | Everyone is good for the country. | I don't like a coward and I don't think it is good for the country. | |
| 2 | የሆንሽ ወሸላ ምን ዶላሽ እዚህ ውስጥ . አፉፉቱ ቱልቱላ | እባክህ ያለውን እንደ ማስታወሻ ይቀበሉ | አይ አንቺ ምን እዚህ ውስጥ ከተተሽ | The model halucinats |
| | You are uncircumcised, Why you are here, talkative | Please accept it as a note. | What are you doing here? | |
| 3 | ወሮ በላ፣ የሰው ፈላጊ፡ ለስራ የሰነፈ ለፀብ የፈጠነ፣ በትዕቢት የተወጠረ | ወሮ በላ፣ እውነተኛ የሰው ፈላጊ ስለ ስራ የተሰነፈ እና የፀብ የፈጠነ ነው። | የራሱ ያልሆነን የሚመኝ፣ ስራን የማይወድ፣ ትህትና የሌለው | The model don't detoxify compound words well and also halucinats |
| | Marauder, seeker of others own, lazy for work, quick for altercation, stiff-necked with vanity. | Marauder , a true man, seeking others own, lazy and quick to anger. | He who desires what is not his own, who does not like work, who has no humility | |

Table 3: Comparison of Gpt4 detoxification outputs Vs human expert detoxification results

| | Input Sentence | Negative Connotations | Shap_detoxified |
|---|---|---|---|
| 1 | ድንቁርናን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ኮልኮሌዎች የሚክዱት ፋክት | የወያኔ, ኮልኮሌዎች | አለመወቅን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ደገፊዎች . የሚክዱት እውነት ነው |
| | The fact which TPLF claptraps deny the fact, sticking on their stupidity as knowledge. | TPLF's Claptraps | The fact which TPLF supporters deny the fact, sticking on their ignorance as knowledge. |
| 2 | እኔ በጣም ፈርቻለው ሰውየው በጣም ደንቆሮና ጀብደኛ ነው አገሪቷን ይዞ ገደል ከመግባት ወደ ኋላ የሚል ሁሉ አይመስልም | አገሪቷን, ሁሉ | እኔ ስጋት አለኝ መሪ ሀገሩን በጥበብ የማይመራ ከሆነ ውጤቱ ተያይዞ መውደቅ እንዳይሆን |
| | I am very afraid that the man is very stupid and reckless. He may take the country down the cliff. | The country, all | I am worried that if the leader does not lead the country wisely, the result will be a collective failure. |
| 3 | ወሮ በላ፣ የሰው ፈላጊ፡ ለስራ የሰነፈ ለፀብ የፈጠነ፣ በትዕቢት የተወጠረ | የሰነፈ, ወሮ | የራሱ ያልሆነን የሚመኝ፣ ስራን የማይወድ፣ ትህትና የሌለው |
| | Marauder, seeker of others own, lazy for work, quick for altercation, stiff-necked with vanity. | Lazy, invade | He who aspires what is not his own, who does not like work, who has no humility. |
| 4 | ልጁ ጅል ቢጤ ስለሆነ ትንኮሳ ነው ብለህ እለፈው?? | ጅል, ብለህ | ልጁ ትንሽ የዋህ ስለሆነ እንደ ትንኮሳ ልንወስደው እንችላለን |
| | Since the boy is a buffoon, just leave it, as it is provocation. | Buffoon, suppose | The boy is a bit naïve, we can take it as provocation. |
| 5 | ብቻህን ሮጠህ ብቻህን አሸናፊ የዘመናችንአስነዋሪ ሰው የዘምናችን ውሻ ቆሻሻ ሰው ነህ | ውሻ, ቆሻሻ | በዚህ ዘመን ብቻውን ሮጦ ብቻውን የሚያሸንፍ አለ ብሎ ለመቀበል አይቻልም:: |
| | You run alone and winer of yourself, you are immodest, dirty, ugly man of our time. | Ugly, dirty | In this day, it is impossible to accept that anyone can run alone and win alone. |

Table 4: Shap outputs of toxic words and detoxified examples.

| | Input Sentence | Negative Connotations | Gpt4_detoxified |
|---|---|---|---|
| 1 | ድንቁርናን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ኮልኮሌዎች የሚክዱት ፋክት | ድንቁርናን, ኮልኮሌዎች | አለመወቅን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ደገፊዎች . የሚክዱት እውነት ነው |
| | The fact which TPLF claptraps deny the fact, sticking on their stupidity as knowledge. | stupidity, claptraps | The fact which TPLF supporters deny the fact, sticking on their ignorance as knowledge |
| 2 | እኔ በጣም ፈርቻለው ሰውየው በጣም ደንቆሮና ጀብደኛ ነው አገሪቷን ይዞ ገደል ከመግባት ወደ ኋላ የሚል ሁሉ አይመስልም | ደንቆሮና, ጀብደኛ | እኔ ስጋት አለኝ መሪ ሀገሩን በጥበብ የማይመራ ከሆነ ውጤቱ ተያይዞ መውደቅ እንዳይሆን |
| | I am very afraid that the man is very stupid and reckless. He may take the country down the cliff. | Stupid, reckless | I am worried that if the leader does not lead the country wisely, the result will be a collective failure. |
| 3 | ወሮ በላ፣ የሰው ፈላጊ፡ ለስራ የሰነፈ ለፀብ የፈጠነ፣ በትዕቢት የተወጠረ | ፈላጊ, የተወጠረ | የራሱ ያልሆነን የሚመኝ፣ ስራን የማይወድ፣ ትህትና የሌለው |
| | Marauder, seeker of others own, lazy for work, quick for altercation, stiff-necked with vanity. | seeker, Stiff-necked | He who aspires what is not his own, who does not like work, who has no humility. |
| 4 | ልጁ ጅል ቢጤ ስለሆነ ትንኮሳ ነው ብለህ እለፈው?? | ጅል, ትንኮሳ | ልጁ ትንሽ የዋህ ስለሆነ እንደ ትንኮሳ ልንወስደው እንችላለን |
| | Since the boy is a buffoon, just leave it, as it is provocation. | Buffoon, Provocation | The boy is a bit naïve, we can take it as provocation. |
| 5 | ብቻህን ሮጠህ ብቻህን አሸናፊ የዘመናችንአስነዋሪ ሰው የዘምናችን ውሻ ቆሻሻ ሰው ነህ | አሸናፊ, ቆሻሻ | በዚህ ዘመን ብቻውን ሮጦ ብቻውን የሚያሸንፍ አለ ብሎ ለመቀበል አይቻልም:: |
| | You run alone and winer of yourself, you are immodest, dirty, ugly man of our time. | Winner, Dirty | In this day, it is impossible to accept that anyone can run alone and win alone. |

Table 5: Shap outputs of toxic words and detoxified examples.

## Conclusion and Future Works

- Presented a new detoxification dataset for Amharic
- Conducted 3 tasks using transformer models: classification (detoxifiable or not), explainability (why a message is toxic) and detoxifiability (rewriting messages).
- GPT4 is better in Amharic text detoxification task despite its frequent hallucination challenges.
- LLMs still struggle with detoxifying low-resource language messages, needs fine-tuning with more datasets